Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

Título: Modelización GLM en el seguro de automóvil

Autoría: Laura Blanco Tomás

Tutoría: Miguel Angel Santolino Prieto

Curso académico: 2020-2021

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Final de Máster

Máster en Ciencias Actuariales y Financieras

Modelización GLM en el seguro de automóvil

Autoría: Laura Blanco Tomás

Tutoría: Miguel Angel Santolino Prieto

El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

Modelización GLM en el seguro de automóvil

Laura Blanco Tomás

Resumen

El presente trabajo se centra en un análisis multivariante de la frecuencia siniestral sobre la base de datos "Singapore Automobile Claims" ofrecida por Frees (2010) para una cartera de seguros de automóvil. Dicho análisis se realizará con modelos GLM, concretamente el modelo de regresión de Poisson, el modelo de regresión Binominal Negativa, el modelo de regresión de Poisson cero inflado y el modelo de regresión Binomial Negativa cero inflado. Aplicados los modelos se identifica cómo afectan los factores de riesgo en la frecuencia siniestral y cuál de ellos se ajusta mejor a los datos.

Palabras clave: siniestros, seguro de automóvil, factores de riesgo, regresión de Poisson, regresión Binomial Negativa.

Abstract

This master's thesis is focused on a multivariate analysis of the loss frequency based on the database "Singapore Automobile Claims" provided by Frees (2010) for a portfolio of automobile insurance. This database is analyzed by means of GLM models, specifically the Poisson regression model, the Negative Binominal regression model, the Inflated Zero Poisson regression model and the Inflated Zero Negative Binomial regression model. Applying these models, it is possible to identify how risk factors impact on the loss frequency and which of them best fits the data.

Key words: claims, automobile insurance, risk factors, Poisson regression model, Negative Binomial regression model.

INDICE DE TABLAS

INDICE DE FIGURAS

ÍNDICE

1. Introducción

Para poder gestionar una cartera de seguros de automóvil resulta de vital importancia controlar los indicadores fundamentales del negocio para garantizar su rentabilidad. En lo que se refiere a los ingresos de la compañía aseguradora resulta imprescindible asegurar una prima media que garantice la cobertura del coste total de los siniestros, así como los gastos, comisiones y beneficios de la compañía.

El coste total de los siniestros constituye la parte más significativa a cubrir por la prima en la mayoría de los productos, pudiendo existir alguna excepción en ramos minoritarios de poca siniestralidad, por ello resulta primordial entender su comportamiento y resultado. Dicho coste total de los siniestros se construye a partir de la frecuencia siniestral, que indicará el volumen de siniestros acontecidos en un determinado periodo temporal, y del coste medio de los mismos.

En el presente trabajo se decide abordar el análisis de la frecuencia siniestral pues presenta variabilidad e incertidumbre ya que el coste medio tiene una vinculación más directa con el coste del objeto asegurado, los costes de las piezas de sustitución, el coste de la mano de obra o las cuantías pactadas por convenios siniestrales entre compañías (por ejemplo, los convenios CIDE o ASCIDE en automóviles).

Por todo ello, el presente trabajo tiene el objetivo de analizar e investigar el comportamiento de la frecuencia siniestral y la importancia que presentan los factores de riesgo de las pólizas en una cartera de automóviles.

1.1. Objetivos

La comprensión de la incidencia de los distintos factores de riesgo sobre la frecuencia siniestral es imprescindible para la construcción de la tarificación de riesgos a priori. Será importante establecer una tarifa a priori muy impactada por dichos factores de riesgo para evitar la antiselección en la contratación de pólizas, dado que los competidores estarán segmentando los riesgos. El hecho de no construir una tarifa suficientemente diferenciada para los distintos factores de riesgo podría provocar que la compañía captase una cuota de mercado significativa de riesgos no deseables perdiendo la rentabilidad del portafolio.

Otro aspecto importante por considerar es que en el mundo asegurador actual existen múltiples formas de negocio que presentan limitaciones en la recopilación de factores de riesgo debido al modelo de negocio. Un ejemplo de ellos es la comercialización de seguros en el momento de la compra de un vehículo o en los modelos de renting. En estos casos los factores de riesgo a solicitar al cliente están muy limitados pudiendo emplear un número muy limitado de factores de riesgo. Conocer qué factores son los más determinantes en la frecuencia siniestral permitirá a la compañía realizar una segmentación del riesgo más precisa pudiendo acordar con mayor facilidad con la empresa automovilística o la empresa de renting los factores de riesgo a incluir en su segmentación de riesgos.

No obstante, comprender la incidencia de los factores de riesgo sobre la frecuencia siniestral será importante para las compañías aseguradoras no solo para definir sus tarifas sino también establecer criterios de suscripción específicos. Por ejemplo, una compañía podría rechazar la suscripción de un riesgo en caso de tener riesgos que incluyan múltiples

factores de riesgo agravados ya que esto evitará de forma directa la entrada de riesgos no deseables para la compañía.

Adicionalmente también es posible emplear estos factores de riesgo en pólizas ya existentes en la cartera pudiendo incorporar alguna de estas variables en los modelos de renovación de pólizas de la compañía o bien estableciendo criterios de saneamiento más restrictivos en riesgos con frecuencias siniestrales agravadas.

Por tanto, mencionados los puntos anteriores, se considera imprescindible comprender el impacto de los factores de riesgo en la frecuencia siniestral para poder garantizar la rentabilidad de la cartera.

1.2. Estructura del trabajo

Para poder realizar este análisis se emplearán los Modelos Lineales Generalizados (GLM) para el recuento de siniestros. Por ello, en la primera parte del trabajo se presentan dichos modelos, explicando los componentes del modelo, la estimación de sus parámetros y los distintos modelos GLM empleados: el modelo de regresión de Poisson, el modelo de regresión Binomial Negativa, el modelo de regresión de Poisson cero inflado y el modelo de regresión de Binomial Negativa cero inflado.

En segundo lugar, se realizará el tratamiento y análisis de la base de datos. Los datos analizados en el presente trabajo recogen las pólizas y siniestros de una cartera de automóviles de Singapur ofrecida por Frees (2010). Esta base de datos ha requerido analizar los valores missing teniendo que tratar y simular algunos datos.

Una vez que se adecuan los datos se realiza un análisis descriptivo de los distintos factores de riesgo recogidos en la base de datos y de su comportamiento siniestral en función de las distintas opciones existentes en cada factor.

También se realiza un análisis de sobre dispersión de los datos. Entre los distintos modelos utilizados se emplean modelos que suponen distintos grados de dispersión entre sus premisas de cálculo por lo que se necesita entender si los datos presentan sobredispersión o no para entender en qué medida estos modelos son acordes a los datos existentes.

En tercer lugar, se procede a aplicar los cuatro modelos anteriormente mencionados analizando e interpretando los outputs obtenidos de dichos modelos. En este caso, es posible llegar a entender de qué forma impactan los distintos factores de riesgo entendiendo como evoluciona la frecuencia siniestral en función del riesgo asegurado. Estos modelos recogen modelos cero inflados que serán de gran relevancia en carteras de seguros pues es habitual encontrar un volumen significativo de pólizas sin siniestros.

Tras la aplicación y análisis de dichos modelos se procede a la selección del modelo que mejor se ajusta a los datos en función de un análisis de bondad de ajuste de cada modelo y de los criterios de información Akaike (AIC) y Bayesiano (BIC).

Todos estos análisis se realizan a través del software estadístico R Studio por lo que se incluye el código en los anexos del trabajo.

2. Metodología

2.1. La construcción de la prima

Para la construcción de una tarifa para el seguro de auto se hace importante el uso de la clasificación de riesgos.

Para cubrir los riesgos que el asegurado ha cedido a la compañía será necesario que la aseguradora cuantifique el importe de la prima pura o prima de riesgo. La prima pura o prima de riesgo es la prima necesaria para poder cubrir los riesgos asegurados. Por tanto, estaría constituida por las pérdidas y gastos esperados. Adicionalmente, habría que incluir el beneficio a obtener.

Una vez obtenida la prima pura la compañía deberá incorporar otros conceptos hasta alcanzar la prima recibo que deberá cobrar al cliente (Sarrasí, 2019):

- Prima de inventario= prima pura + gastos de gestión interna.
- Prima de tarifa= prima de inventario + los gastos de gestión externa (donde destacan las comisiones).
- Prima recibo= prima de tarifa + impuestos + consorcio + otros

Dado que los riesgos captados por la aseguradora son en la mayoría de los casos heterogéneos, es de vital importancia realizar una tarifa sofisticada que genere distintas primas en función de los distintos perfiles de riesgo. Cuanto mayor sea el conocimiento de las características del riesgo asegurado, también conocidas como factores de riesgo o tarificación, más precisa será la segmentación de las tarifas.

2.2. Los factores de tarificación

El mercado asegurador presenta cada vez tarifas más sofisticadas y las compañías tienen en cuenta las características del riesgo o póliza para la construcción de primas. Los riesgos que presenten una reducida probabilidad de ocurrencia de siniestros presentarán una prima media inferior en el mercado. De igual forma, los riesgos con elevadas probabilidades de presentar siniestros tendrán una prima media superior en el mercado en busca de cubrir con las primas el riesgo asegurado.

De esta forma, si una compañía no tiene en cuenta los factores de riesgo ofrecerá la misma prima a todos los riesgos. De esta forma, captará los riesgos con mayor probabilidad siniestral ya que el resto de los competidores los habrán identificado y les ofrecerán primas superiores. De la misma forma, los riesgos con menor probabilidad siniestral serán captados por los competidores ya que ofrecerán primas más competitivas.

Por tanto, gracias a la inclusión de los factores de riesgo en la tarifa la compañía podrá evitar la antiselección o selección adversa de riesgos.

Gracias a un sistema de clasificación la aseguradora creará clases de riesgo y podrá asignar cada riesgo o póliza a una clase de riesgo en base de sus factores de riesgo, también conocidos como factores de tarificación. Los riesgos o pólizas con factores de riesgo iguales se asignarán a una misma clase siendo esta dentro de si misma homogénea pero heterogénea respecto al resto de clases.

Los factores de riesgo o factores de tarificación recogen las características del riesgo y son variables a priori, pues serán conocidas para la aseguradora antes de la cesión del riesgo.

Lo anteriormente comentado hace primordial la identificación y selección de aquellas características del riesgo que estén relacionadas con el resultado siniestral de las pólizas, siempre que no haya algún impedimento legal para su tratamiento.

De esta forma, se generará una relación entre los factores de tarificación y el resultado siniestral por lo que la aseguradora buscará estimar la frecuencia siniestral a partir de los factores de riesgo. En este caso, la aseguradora hará uso de modelos de regresión lineales generalizados en los que introducir los factores de clasificación que le permita estimar unos resultados siniestrales a partir de ellos. No obstante, podrían aplicarse otras metodologías como por ejemplo modelos de regresión lineal múltiples o modelos de clasificación de riesgos (árboles de decisión).

Uno de los modelos que permiten realizar dicha clasificación de riesgos es el modelo de Poisson cuya medición de siniestralidad está focalizada en la estimación de la frecuencia siniestral (número de siniestro).

2.3. Modelos lineales generalizados

El objetivo del presente trabajo será aplicar modelos lineales generalizados con el objetivo de conocer de qué forma los distintos factores de riesgo afectan al número de siniestros. En este caso, el número de siniestros constituirán la variable dependiente endógena o variable objetivo que vendrá explicada por las distintas variables explicativas exógenas o independientes constituidas por los distintos factores de riesgo o factores de tarificación.

Se aplicarán modelos lineales generalizados ya que para poder modelizar el número de siniestros no será posible aplicar un modelo de regresión clásico ya que:

- Se desea modelizar el número de siniestros cuya variable aleatoria es discreta. El modelo de regresión clásico supone normalidad analizando variables continuas. Adicionalmente, en la modelización del número de siniestros, frecuentemente hay asimetría a la derecha. Por su parte, los modelos lineales generalizados permitirán aplicar una clase general de distribuciones de la familia exponencial pudiendo aplicar otras distribuciones distintas a la normal.
- La regresión lineal clásica supone homocedasticidad. En cambio, es de esperar que la desviación estándar del número de siniestros no se mantenga siempre constante para distintos valores de las variables explicativas.
- En los modelos lineales generalizados la esperanza de la variable dependiente se relacionará con las variables independientes linealmente pero no de forma directa sino a través de la función enlace.

2.3.1. Componentes del modelo

Se diferencian tres componentes en los modelos lineales generalizados: la componente aleatoria, la componente sistemática y la función enlace (Alemany, Ayuso y Bolancé, 2021).

2.3.1.1. Componente aleatoria

Se dispondrá de Y_i como variable objetivo-endógena para $i = 1, \dots, n$ pólizas teniendo $Y_1, \dots, Y_n$.

La variable Y constituirá la componente aleatoria y seguirá una distribución de la familia exponencial. En este caso, dado que el número de siniestro abarca valores enteros positivos se emplearán para el presente trabajo las distribuciones de Poisson y Binomial Negativa.

2.3.1.2. Componente sistemática

La componente sistemática también conocida como predictor lineal se define a través de las $j = 1, \dots k$ variables independientes exógenas explicativas para $i = 1, \dots, n$ pólizas $X_{i1}, \dots, X_{ik}$ y sus coeficientes a estimar $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$.

$$\eta_i = \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}\beta_j = X'\beta = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \quad \forall i = 1, \dots, n$$

A la hora de modelizar los datos de la cartera se tendrán en cuenta los distintos factores de riesgo asociados a la póliza o riesgo. Se deberá relacionar el valor esperado de siniestros con los distintos factores de riesgo. De esta forma se obtendrán parámetros de frecuencia cambiantes en función del nivel de riesgo de los distintos factores de riesgo.

En caso los factores de riesgo pueden hacer referencia a variables categóricas con distintas categorías de respuesta (por ejemplo, tipo de vehículo turismo, furgoneta o motocicleta), cada categoría constituirá un nivel (h) para un factor de riesgo y estos niveles serán numerados de forma arbitraria ($h = 1,2, \dots, l$).

Se podrá, por lo tanto, modelar l niveles representados por $l - 1$ variables calculando la siguiente media:

$$\eta_i = \sum_{i=1}^{n} \sum_{j=1}^{l-1} X_{ij}\beta_j = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_{l-1} X_{il-1}, \quad \forall i = 1, \dots, n$$

2.3.1.3. Función enlace

La función enlace $g(\mu_i)$ permitirá relacionar el valor esperado de siniestros o la variable dependiente con el predictor lineal de las variables explicativas:

$$g(\mu_i) = \eta_i = X_i'\beta$$

Dado que se desea modelizar el número de siniestros se necesitarán modelos para conteo, haciendo uso de los modelos de Poisson y Binomial Negativa, para ambas funciones la función enlace será la función logarítmica.

$$g(\mu_i) = log(\mu_i) = \eta_i = X_i'\beta \qquad \mu_i = e^{X_i'\beta}$$

2.3.2. Estimación de Máxima Verosimilitud

Se estimarán los parámetros del predictor lineal en base al modelo escogido a través del método de Máxima Verosimilitud por Alemany, Ayuso y Bolancé (2021). Con el método de Máxima Verosimilitud se buscarán los valores de los parámetros β.

Partiendo de una función de densidad de la familia exponencial el logaritmo de la función de verosimilitud será para n observaciones independientes:

$$\ln L(y;\theta,\phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \frac{\phi}{\omega_i})$$

Siendo ϕ un parámetro de dispersión el cual incluye un ponderador ω_i, θ_i los parámetros a estimar y las funciones b() y c() dependerán la función de densidad de la familia exponencial cuyo conjunto de distribución se adaptará a esta estructura.

Para conocer los valores de los parámetros se derivará el logaritmo por cada parámetro, siendo l_i el i-ésimo componente del logaritmo:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} x \frac{\partial \theta_i}{\partial \mu_i} x \frac{\partial \mu_i}{\partial \eta_i} x \frac{\partial \eta_i}{\partial \beta_j}, \quad \forall j = 1, \dots, k$$

Para poder obtener dichos parámetros se deberá trabajar con aproximaciones obtenidas a través de procedimientos interactivos tales como el algoritmo de Newton-Raphson o el método Scoring de Fisher.

2.3.3. Modelos de recuento de siniestros

Ya que en este caso la variable dependiente hace relación al número de siniestros tomando valores discretos positivos, será necesario aplicar un modelo de recuento de siniestros. El número de siniestros será la variable dependiente endógena o variable objetivo y habrá una serie de variables explicativas exógena o independientes constituidas por los distintos factores de riesgo o factores de tarificación que explicarán el comportamiento de la variable dependiente.

Los modelos propuestos para representar esto datos será:

- El modelo de regresión de Poisson
- El modelo de regresión Binomial Negativa
- El modelo de regresión de Poisson cero inflado
- El modelo de regresión de Binomial Negativa cero inflado

2.3.3.1. Regresión de Poisson

Tal y como detalle Frees (2020) en el libro Loss Data Analytics. (pp.285- 308), siendo Y la variable dependiente a analizar, la función de densidad de una Poisson de media y varianza μ es:

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0,1,2, \dots$$

siendo $E(Y) = Var(Y) = \mu$.

A la hora de modelizar los datos de la cartera con la regresión de Poisson se tendrán en cuenta los distintos factores de riesgo asociados a la póliza o riesgo con l niveles representados por $l-1$ variables calculando la siguiente media:

$$\log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{l-1} x_{l-1}$$

Para evitar el impacto de valores negativos se puede emplear la relación lineal en lugar del log lineal. Supongamos que únicamente tenemos como variable explicativa la variable categórica con l niveles, calculándose para $z = 1, 2, \dots, l-1$:

$$\mu = e^{\beta_0 + \beta_z}$$

Donde x_z tomará valores 1 y 0:

$$x_z = \begin{cases} 1 \ \ para\ nivel\ z \\ 0 \ \ para\ nivel\ distinto\ a\ z \end{cases}$$

De esta forma se irán obteniendo valores de 0 y 1 para las distintas $l-1$ categorías del factor de riesgo.

De esta forma el nivel base será el nivel l, obteniendo la media siguiente en términos líneas y no logs lineales:

$$\mu = \begin{cases} e^{\beta_0 + \beta_z} & (nivel\ z = 1,2, \dots, l-1) \\ e^{\beta_0} & (nivel\ z = l) \end{cases}$$

Relacionando el parámetro μ para lo distintos niveles $z = 1, 2, \dots, l-1$:

$$\mu_{(z)} = \mu_{(l)} e^{\beta_z}$$

Siendo e^{β_z} será un factor corrector que aplicar sobre valor de base definido en el nivel l con $\mu_{(l)}$ para cualquier frecuencia esperada para los niveles $z = 1,2, \dots, l-1$.

Disponiendo de una cartera de n pólizas independientes con un conjunto de factores de tarificación en forma de vector de dimensión k constituido por sus distintos niveles.

El factor de tarifación para la póliza $i-ésima$ constituiría el siguiente vector:

$$x_i = (1, x_{i1}, \dots, x_{ik})' \qquad \text{para } i = 1, \dots, n$$

Con una siniestralidad asociada a cada póliza i:

$$y_i \in \{0,1,2, \dots\}$$

Estos factores de tarificación x_i son las variables independientes explicativas del modelo que ofrecerán información acerca de la variable dependiente que serán la frecuencia siniestral y_i.

La media de la regresión de Poisson que representará la frecuencia esperada para la póliza $i-ésima$ tomando la función exponencial que permitirá que $\mu_i > 0$:

$$\mu i = E\,(y_i | x_i) = \ e^{x_i' \beta}$$

O en su forma lineal:

$$\log \mu i = \log E\,(y_i|x_i) = \; x_i'\beta$$

$$y_i \sim Pois(\mu_1), i = 1, \ldots, n$$

De forma que dado que tenemos los coeficientes $\beta = (\beta_0, \ldots, \beta_k)'$, se obtendría:

$$x_i'\beta = \beta_0 + \beta_1\, x_{i1}, \ldots, \beta_k x_{ik}$$

Las fórmulas hasta ahora introducidas para la regresión de Poisson consideran que la **exposición del riesgo** durante el periodo analizado (un año) será completa y que, por tanto, todas las pólizas tendrán una exposición de 1.

Para introducir la exposición incluiremos el parámetro t de la pmf de Poisson en términos de parámetro λ que indicará el número de siniestros. La función de distribución de Poisson resultante es:

$\Pr(Y = y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!}$ con $y = 0,1,2, \ldots$

Con $E(Y) = Var(Y) = \; \lambda t$.

El parámetro λ señalará el número de siniestros por exposición y t señalará la exposición constante y conocida. De esta forma, con el parámetro λ a mayor exposición se producirá un mayor conteo de siniestros.

Disponiendo de λ_i para indicará el número de siniestros (tasa del evento pérdida) y t_i para la exposición y μ_i para el valor esperado de siniestros bajo la exposición existente.

$$\mu i = E\,(y_i|x_i) = \; t_i\lambda_i = t_i e^{x_i'\beta}$$

$$y_i \sim Pois(\mu_i), i = 1, \ldots, n$$

$$\log \mu i = \log t_i + x_i'\beta$$

$$y_i \sim Pois(\mu_i), i = 1, \ldots, n$$

De esta forma, $\log m_i$ también llamado offset constituirá una variable explicativa del modelo adicional que podremos incluir al tratarse de una constante a la que, en este caso, se le aplicará el coeficiente 1.

2.3.3.2. Regresión de Binomial Negativa

La regresión binomial negativa se basa en una distribución que mezcla Poisson y Gamma. Ya no parte de la premisa restrictiva de Poisson de que la varianza deberá ser igual a la media y esto permite modelar la heterogeneidad de a través de la distribución gamma.

La regresión de Poisson presenta la propiedad de equidispersión e los datos, por tanto, en caso de que los datos contengan sobredispersión no podrá ser capturada por el modelo de Poisson y esto puede provocar que se subestimen los errores de los coeficientes pudiendo determinar un factor de riesgo como significativo que realmente no lo sea. El modelo de regresión binomial negativo, en cambio, podrá captar parte de la varianza que no está captando la Poisson y pudiendo realizar un mejor ajuste.

En la muestra actual de datos existen factores de riesgo no contemplados asociados al objeto asegurado o al conductor del vehículo, algunos ejemplos pueden ser la experiencia del conductor o la ubicación geográfica del riesgo. Estos factores de riesgo no recogidos es muy probable que estén teniendo impacto en la frecuencia siniestral y que generen variabilidad provocando una heterogeneidad no observada. Esta heterogeneidad estaría generando sobredispersión en el modelo de Poisson ya que establece que los siniestros son independientes y aleatorios entre los distintos individuos. Nos encontraríamos que la varianza observada sea superior a la varianza teórica, que en Poisson será igual a la media.

La media de la regresión de Binomial negativa que representará la frecuencia esperada para la póliza $i-ésima$ tomando la función exponencial que permitirá que $\mu_i > 0$ incluyendo un factor de heterogeneidad ε_i o perturbación en la media de Poisson con una distribución Gamma:

$$\mu i = E\,(y_i|x_i) = \; e^{x_i'\beta+\varepsilon_i}$$

O en su forma lineal:

$$\log \mu i = \log E\,(y_i|x_i) = \; x_i'\beta + \varepsilon_i$$

Dicha heterogeneidad puede recogerse a través de la regresión binomial negativa. La función de densidad de la binomial negativa será la siguiente (NCSS Statistical Software, p. 326-1):

$$\Pr(Y = y) = \frac{\Gamma(\mathrm{y} + \alpha^{-1})}{\Gamma(\mathrm{y}+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}$$

Donde:

$$\mu = t\mu$$

$$\alpha = \frac{1}{v}$$

Siendo μ la media por unidad expuesta t en este caso como sucedía con la regresión de Poisson medida a partir del tiempo durante el cual la póliza se ha encontrado expuesta al riesgo. El parámetro v será un parámetro de escala y el parámetro α constituirá un parámetro de dispersión presentando equidispersión, por tanto, en caso de que sea igual a 0.

La regresión binomial negativa modelizará los datos como una regresión de Poisson, pero contemplará la heterogeneidad individual no observada a través de la inclusión de la distribución gamma, contemplando que únicamente con la media no se recoge toda la información.

De la misma forma que sucedía con la Poisson podremos modelar l niveles representados por $l-1$ variables calculando la siguiente media:

$$\log \mu = \log t_i + \beta_0 + \beta_1 x_1 + \cdots + \beta_{l-1} x_{l-1}$$

2.3.3.3. Modelos de regresión cero inflados

La regresión Binomial Negativa recoge en mayor medida la sobredispersión que la regresión de Poisson que contempla equidispersión donde la media siniestral es igual a la varianza.

Sin embargo, en una base de datos como la que se va a tratar en el presente trabajo nos podemos encontrar con una sobredispersión de los datos significativa dado que la varianza sería significativamente superior a la media motivada por la presencia de un volumen importante de pólizas sin siniestros.

Esta presencia importante de ceros provocará que las distribuciones de Poisson y Binomial Negativa puedan subestimar el error de las variables explicativas identificando que son significativas pudiendo no serlo. Por tanto, la precisión de ambos modelos ser verá afectada y será necesario acudir a modelos distintos que corrijan este problema.

Para poder trabajar con una base de datos con un volumen tan significativo de ceros se plantea (López, 2019):

- Modelos truncados: en los modelos truncados se partirá la base de datos en dos partes, una de ellas con ceros y la otra sin ceros. De esta forma primero se modelizará la probabilidad de obtener un cero y posteriormente se aplicará una distribución truncada en cero.

- Modelos cero inflados: en los modelos cero inflados se contempla la existencia de dos fuentes distintas en la obtención de ceros, una de ellas procedentes por un uso reducido del riesgo asegurado en la póliza y el segundo de ellos porque pese a un uso significativo del riesgo no se producen siniestros. Esto se traducirá en que en la primera parte del proceso de modelizará la probabilidad de obtener ceros procedentes del reducido uso del riesgo asegurado y en segundo lugar se modeliza la probabilidad de realizando uso del riesgo obtener cero o más siniestros.

En nuestra base de datos recogemos pólizas de la línea de automóviles. En una base de datos de pólizas de automóviles es habitual encontrar un volumen importante de ceros debido a que habrá una parte de conductores que no hayan utilizado mucho el vehículo y que por tanto su frecuencia siniestral tenderá a cero, y la segunda de ellas vendrá motivada por una parte de conductores del portafolio con buenos hábitos de conducción.

Para poder realizar nuestra modelización de la frecuencia siniestral se optará por el uso de modelos cero inflados ya que como se ha mencionado realizarán un primer análisis identificando y midiendo el volumen de ceros procedentes del bajo uso de los vehículos, y posteriormente habiendo identificado ya esa casuística procederá a modelizar el resto de los valores incluyendo los ceros procedentes de hábitos de conducción. Estos modelos contemplarían la premisa anteriormente comentada de que los ceros procederán de varias fuentes. Este hecho descarta los modelos truncados.

Los modelos truncados también están constituidos por dos procesos, no obstante, en el primero de ellos realizarán una cuantificación de valores cero y en una segunda parte del proceso identificarán valores distintos a cero. Esto hará que estos modelos no sean adecuados para nuestro caso ya que pese a que también constan de dos procesos identificarán que los ceros procederán únicamente de la primera componente ya que en su segunda componente aplica modelos truncados en cero que no admitirán la existencia de ceros.

Dentro de los modelos cero inflados nos podremos encontrar con:

- Modelo de regresión de Poisson cero inflado: como ya se ha comentado al tratarse de un modelo de cero inflados estará formado por dos procesos, el primero de ellos modelizará probabilidad de valores de cero y en segundo lugar aplicará el modelo de Poisson.
  En aras de aplicar este modelo se identificará a través de $p_i$la probabilidad de que la póliza i-ésima presente cero siniestros y una probabilidad $1 - p_i$ que sus valores procedan de una distribución de Poisson que recogerá valores iguales o superiores a cero.

  Su función de probabilidad vendrá dada por:

$$\Pr(Y_i = 0) = p_i + (1 - p_i)e^{-\lambda_i}$$
$$\Pr(Y_i = k) = \frac{(1 + p_i)e^{-\lambda_i}\lambda_i^k}{k!}, \qquad siendo\ k > 0$$

  Donde los vectores de λ_i y p_i:

$$\log(\lambda_i) = Z_i'\beta$$
$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = W_i'\gamma$$

  siendo Z_i y W_i vectores de variables explicativas, y sus coeficientes a estimar β y γ.

- Modelo de regresión de Binomial Negativa cero inflado: como en el caso anterior al tratarse de un modelo de cero inflados estará formado por dos procesos, el primero de ellos modelizará probabilidad de valores de cero y en segundo lugar aplicará el modelo de Binomial Negativo.

  Su función de probabilidad vendrá dada por:

$$\Pr(Y_i = 0) = p_i + (1 - p_i)\left(\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}}\right)^{\alpha^{-1}}$$
$$\Pr(Y_i = k) = (1 - p_i)\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(1 + y_i)\Gamma(\alpha^{-1})}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}},$$
$$siendo\ k > 0$$

3. Base de datos

Para el presente trabajo se empleará la base de datos "Singapore Automobile Claims" que se encuentra disponible en el Manual para modelos de regresión con aplicaciones actuariales y financieras publicado por Frees (2010) cuya información procede de la Asociación General de Seguros de Singapur (GIA) y recoge el volumen de siniestros para un conjunto de pólizas del seguro de automóvil en Singapur, así como varios factores de riesgo.

Adicionalmente, se encuentra disponible en el paquete de R 'insuranceData'. Este paquete recoge bases de datos que permiten trabajar en la modelización de la frecuencia y sus autores son Michal Trzesiok y Alicja Wolny--Dominiak.

3.1. Presentación de los datos

Los variables que aparecen en la base de datos analizada son los siguientes:

Tamaño de la muestra	7,483
Obs. con valores missing	3,641

Variable	*Descripción*
SexInsured	Genero del asegurado. M → Hombre F → Mujer U → No especificado
Female	1 → Mujer 0 → Otros
VehicleType	Tipos de vehículo A, G, M, P, Q, S, T, W, Z
PC	1 → Uso privado 0 → Otros
Clm_Count	Numero de siniestros durante el periodo
Exp_weights	Periodo de exposición anual
LNWEIGHT	Logaritmo de la exposición
NCD	Bonificación por no siniestros
AgeCat	Edad del asegurado: 0 → < 21 2 → 22-25 3 → 26-35 4 → 36-45
VAgeCat	Antigüedad del vehículo: 0 → 0 1 → 1 2 → 2 3 → 3-5 4 → 6-10 5 → 11-15 6 → ≥16
AutoAge0	1 → Uso Privado + Antigüedad del vehículo 0 0 → Otros
AutoAge1	1 → Uso Privado + Antigüedad del vehículo 1 0 → Otros
AutoAge2	1 → Uso Privado + Antigüedad del vehículo 2 0 → Otros
AutoAge	1 → Uso Privado + Antigüedad del vehículo 0,1 o 2 0 → Otros
VAgecat1	VAgeCat con categorías 0, 1 y 2 combinadas

Tabla 1. Descripción de variables de la base de datos inicial. Fuente: Elaboración propia.

Será necesario analizar las variables y observaciones de la base de datos con el objetivo de identificar si será necesario tratarla antes de aplicar los modelos a trabajar.

3.2. Tratamiento de los datos

3.2.1. Análisis de valores missing

El tamaño de la muestra es de 7.483 observaciones. No obstante, al analizar la base de datos se puede observar que los tipos de vehículos distintos a Automóvil (A) presentan valores missing en género y valores de 0 en edad y uso del vehículo.

Si bien podría tratarse de un valor real de uso y edad (uso no privado para edades ≤ 21), los volúmenes tan elevados de un mismo valor único ligados a un volumen significativo de género desconocido hacen pensar que dichas observaciones presentan missings.

Tipo Vehículo ≠ A	Nº Observaciones	3641
Género		
U	3638	99.9%
F	0	0.0%
M	3	0.1%
Uso		
0	3641	100.0%
1	0	0.0%
Edad		
0	3640	100.0%
4	1	0.0%

Tabla 2. Identificación valores missing. Fuente: Elaboración propia.

Para poder realizar el trabajo con un volumen importante de observaciones y poder trabajar con los datos se ha procedido a tratar y simular estos factores de riesgo para estas 3641 observaciones. Para simular los campos se han tenido en cuenta:

- Las frecuencias del sector (por ejemplo, género masculino con mayor frecuencia que el género femenino).
- Coherencia con el resto de los factores de riesgo (por ejemplo, asociadas edades menores a bonificaciones inferiores debido a la falta de experiencia que permita conseguir máximas bonificaciones)

3.2.2. Valor exposición

El concepto de exposición será determinante a la hora de calcular la frecuencia siniestral. La exposición hace referencia al tiempo durante el cual el riesgo se encuentra cedido a la aseguradora e indica, con una base de cálculo anual, el periodo durante el cual la póliza está expuesta al riesgo y es susceptible de tener siniestros.

Los asegurados podrán contratar la póliza e iniciar su exposición en cualquier momento del año y, de igual forma, podría darse el caso que de forma excepcional pudiese terminar la exposición antes de terminar su año natural. Este hecho hace que sea imprescindible considerar la exposición para cada póliza y que se tenga en cuenta en la frecuencia siniestral.

En la base de datos existentes se dispone de la exposición por pólizas a través de la variable “Exp_weights” y del número de siniestros por póliza con la variable “Clm_Count”.

Para poder calcular los siniestros en base al tiempo durante el cual la póliza ha estado expuesta, se creará la variable "Exp_Clm". Esta variable será el cociente entre el número de siniestros y el tiempo de exposición y de esta forma obtendremos el número de siniestros en términos anuales para todas las pólizas.

Realizar este cálculo de frecuencia siniestral implicará dar por hecho un comportamiento homogéneo de la siniestralidad a lo largo del año. En este caso, aceptaremos este cálculo como correcto sin realizar otros supuestos adicionales ya que el producto de Auto presenta reducida estacionalidad. En caso de trabajar con otros ramos podría ser conveniente realizar una transformación adicional. Un ejemplo de ellos podría ser el producto de Hogar ya que se encuentra muy expuesto a la meteorología con diferente afectación en función de la estación del año.

3.2.3. Selección de los factores de riesgo

En cuanto a los factores de riesgo a analizar, es importante destacar:

- Las variables AutoAge0, AutoAge1, AutoAge2, AutoAge y VAgecat1 son agrupaciones de variables realizados por los autores de la base de datos. Para este trabajo estas variables serán descartadas ya se partirá de los factores de riesgo originales que podrán ser agrupados posteriormente si se estima necesario.

- La variable Female es una dicotómica de la variable SexInsured que por sí sola no aporta valor adicional. Por este motivo esta variable tampoco se incluirá.

Por tanto, los 6 factores de riesgo a analizar serán:

- El género del tomador (SexInsured)
- Tipo de Vehículo (VehicleType)
- El uso del vehículo (PC)
- El historial del tomador (NCD)
- La edad del tomador (AgeCat)
- La antigüedad del vehículo (VAgeCat)

La base final resultante estará constituida por:

Tamaño de la muestra	7,483
Obs. con valores missing	0

Variable	***Descripción***
SexInsured	Genero del asegurado.
VehicleType	Tipos de vehículo A, G, M, P, Q, S, T, W, Z
PC	1 → Uso privado
Clm_Count	Numero de siniestros del periodo
Exp_weights	Periodo de exposición anual
LNWEIGHT	Logaritmo de la exposición
NCD	Bonificación por no siniestros
AgeCat	Edad del asegurado:
VAgeCat	Antigüedad del vehículo:
Clm_Exp	Numero de siniestros anualizados

Tabla 3. Descripción de variables con la base de datos tratada. Fuente: Elaboración propia.

3.3. Análisis inicial

Tras realizar el tratamiento y homogeneización de los datos se procede a realizar un breve análisis descriptivo de los datos en su globalidad.

Nº de Siniestros	0	1	2	3	Total
Nº Pólizas	6996	455	28	4	7483
% Pólizas	93.5%	6.1%	0.4%	0.1%	100.0%

Tabla 4. Análisis global del número de siniestros declarados para el periodo observado. Fuente: Elaboración propia.

Podemos observar cómo el 93,5% del portafolio no presenta siniestros. Esto suele ser lo habitual pues gracias a los equipos de pricing y suscripción se busca la selección de riesgos preferentes a través de las palancas de la tarifa y del producto o selección de riesgo en el momento de la contratación.

Una buena segmentación de la tarifa a través de los factores de riesgos debería ayudar a aumentar el peso de los riesgos con reducidas frecuencias ayudando a mejorar la frecuencia global.

3.4. Análisis de los factores de riesgo

3.4.1. Género del Tomador

En primer lugar, se empezará por analizar el género del tomador de póliza cuya variable se trata de una variable cualitativa dicotómica.

Se hace importante resaltar que esta variable no podrá ser implementada en el proceso de tarificación, si bien en el presente trabajo se pretende analizar la frecuencia y su comportamiento en función de los factores de riesgo y no tendrá por tanto afectación.

Esto es así debido a la entrada en vigor de la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres. Esta ley prohíbe a las compañías aseguradoras a incluir el género como un factor de tarificación.

El género del tomador es un factor de riesgo que puede tomar 2 valores, masculino o femenino. La distribución de las observaciones para este factor de riesgo es la siguiente:

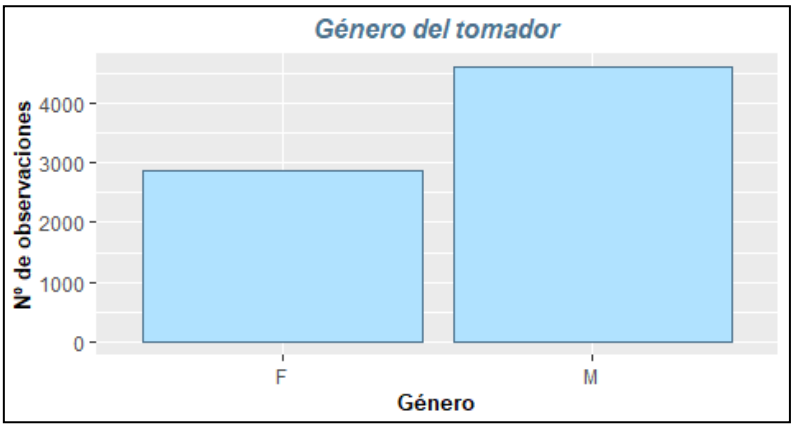

Gráfico 1. Factor de riesgo: Género. Fuente: Elaboración propia.

Tal y como se puede observar en el Gráfico 1, el volumen de hombres en el portafolio es superior con 4.605 observaciones representando un 61,54% del total de la muestra, tal y como se constata en la Tabla 5.

Adicionalmente podemos observar como el género masculino presenta una media de siniestros anuales superior con un 0.16 respecto al valor de 0.10 que presenta el género femenino.

Para el caso de los hombres, estos presentan mayor volatilidad en su frecuencia adicionalmente dado que presentan una mayor desviación estándar. En este caso, la volatilidad también supone mayor incertidumbre y por tanto riesgo en el portafolio.

Adicionalmente podemos ver que los hombres presentan una curtosis superior a las mujeres. Esto hace que la distribución del número de siniestros anuales recoja una mayor probabilidad en su cola donde el número de siniestros es superior. Por tanto, un riesgo masculino tendrá una probabilidad superior de tener un mayor volumen de siniestros por póliza. Esto se refuerza con el cuantil.

	Género		
	Femenino (F)	Masculino (M)	Total
Nº Observaciones	**2,878**	**4,605**	**7,483**
% Obs.	38.46%	61.54%	**100.00%**
Nº Siniestros Anuales	**288**	**732**	**1,020**
Media	0.10	0.16	0.14
Mediana	0.00	0.00	0.00
Desviación estándar	0.58	0.78	0.71
Varianza	0.34	0.60	0.50
Asimetría	11.18	11.41	11.64
Kurtosi	177.03	223.23	229.51
Min	0.00	0.00	0.00
Max	11.78	22.83	22.83
Cuantil .25	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00
Cuantil .95	1.00	1.28	1.17

Tabla 5. Tabla del factor de riesgo género. Fuente: Elaboración propia.

3.4.2. Tipo de Vehículo

Al igual que el género el tipo de vehículo es una variable cualitativa que en este caso consta de un rango de 9 posibles niveles de respuesta: A, G, M, P, Q, S, T, W y Z.

Las descripciones de la base de datos únicamente detallan el significado de tres tipos de vehículo: A como automóvil o turismo si apelamos a las categorías de vehículo comunes (siendo automóvil un término empleado por antonomasia para referirse a los automóviles de turismo), M como motocicleta y T como camión. En este caso,

Atendiendo a la distribución de pólizas existentes que se puede observar en la Tabla 6, podemos ver que no resulta demasiado interesante realizar un análisis de los Camiones (T) por separado ya que su peso es muy reducido. Sin embargo, el tipo de vehículo G resulta de gran interés debido a su volumetría.

	A	G	M	P	Q	S	T	W	Z
Nº Observaciones	3,842	2,882	188	88	358	16	8	30	71

Tabla 6. Tipos de vehículo. Fuente: Elaboración propia.

En Singapur es necesario disponer de un certificado de titularidad (COE) para poder poseer un vehículo (Almoguera y Roslan, 2017). En este caso, en aras de identificar a qué tipo de vehículo está haciendo referencia el tipo de vehículo G se acude a fuentes de información afines al Ministerio de Transportes de Singapur referentes al transporte terrestre (Land Transport Authority). La información publicada por Land Transport Authority arroja que tras los automóviles turismos el tipo de vehículo con mayor volumen es el de vehículos de transporte de bienes. Por tanto, se determina que los tipos de vehículo G harán referencia a Furgonetas.

Realizada esta identificación y dado que el resto de las variables no están identificadas, sólo el tipo de vehículo camión con únicamente 8 pólizas, se decide reagrupar los tipos de vehículo como A, G, M y agrupando todas las demás como O (Otros).

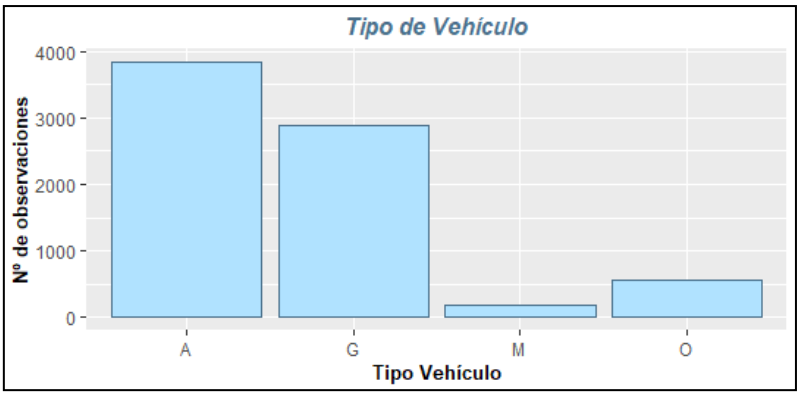

Gráfico 2. Factor de riesgo: Tipo de Vehículo. Fuente: Elaboración propia.

A través del Gráfico 2 se puede observar que la mayoría de los vehículos son o bien turismos (A) o bien furgonetas (G). El resto de vehículo (M y O) presentan un peso únicamente de entorno al 10%, como se observa en la tabla 7.

Adicionalmente, se puede observar que tanto las furgonetas como las motocicletas presentan una media de siniestros inferior a los turismos y otros vehículos, con menor dispersión de la muestra debido además a desviaciones estándar inferiores. Esto puede deberse fundamentalmente a la existencia de restricciones significativas en el proceso de contratación de dichas pólizas. En este caso, se asume que dichas restricciones se mantendrán en para la tarifa en la que se decida utilizar los modelos trabajados en el presente trabajo.

	Tipo de Vehículo				
	Turismos (A)	Furgoneta (G)	Motocicleta (M)	Otros (O)	Total
Nº Observaciones	3,842	2,882	188	571	**7,483**
% Obs.	51.34%	38.51%	2.51%	7.63%	**100.00%**

Nº Siniestros Anuales	**638**	**278**	**5**	**99**	**1,020**
Media	0.17	0.10	0.02	0.17	0.14
Mediana	0.00	0.00	0.00	0.00	0.00
Desviación estándar	0.83	0.48	0.17	0.89	0.71
Varianza	0.68	0.23	0.03	0.79	0.50
Asimetría	11.42	7.39	6.97	9.20	11.64
Kurtosi	212.02	78.81	48.91	105.49	229.51
Min	0.00	0.00	0.00	0.00	0.00
Max	22.83	9.13	1.52	11.78	22.83
Cuantil .25	0.00	0.00	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00	0.00	0.00
Cuantil .95	1.33	1.00	0.00	1.14	1.17

Tabla 7. Tabla del factor de riesgo tipo de vehículo. Fuente: Elaboración propia.

3.4.3. Uso de Vehículo

El factor de riesgo de uso del vehículo es una variable cualitativa dicotómica que diferencia el uso privado o particular del vehículo del uso no privado o no particular. Tal y como se observa en el Gráfico 3 el peso de ambos tipos de usos es muy similar.

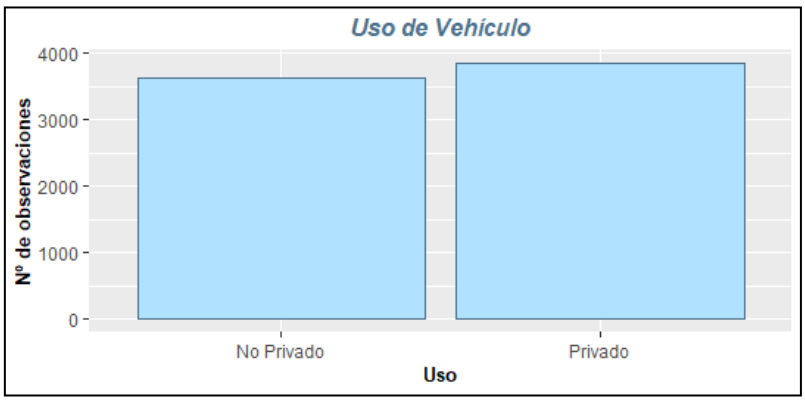

Gráfico 3. Factor de riesgo: Uso de Vehículo. Fuente: Elaboración propia.

Como se puede observar en la Tabla 8 el 51% de los vehículos son de uso privado o particular. Por otra parte, la media de siniestros anuales resulta superior en vehículos particulares que en vehículos no privados.

	Uso del Vehículo		
	Privado (1)	No Privado (0)	Total
Nº Observaciones	**3,851**	**3,632**	**7,483**
% Obs.	51.46%	48.54%	**100.00%**

Nº Siniestros Anuales	**611**	**408**	**1,020**
Media	0.16	0.11	0.14
Mediana	0.00	0.00	0.00
Desviación estándar	0.79	0.61	0.71
Varianza	0.62	0.37	0.50
Asimetría	11.80	10.36	11.64
Kurtosi	235.52	155.62	229.51
Min	0.00	0.00	0.00
Max	22.83	11.78	22.83
Cuantil .25	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00
Cuantil .95	1.24	1.01	1.17

Tabla 8. Tabla del factor de riesgo uso de vehículo. Fuente: Elaboración propia.

En este caso, tal y como se observa en la Tabla 9 esto estará muy ligado al tipo de vehículo:

- Para los turismos o automóviles: la frecuencia aumenta cuando el vehículo no es de uso privado. Esto podría deberse a usos como el taxi que aumentan de forma importante la frecuencia.

- Para los vehículos distintos a turismos: la frecuencia aumenta cuando el vehículo es de uso privado. Esto podría deberse al hecho de que trabajando para un tercero y haciendo uso de su vehículo el empleado y tomador de la póliza sienta una mayor responsabilidad sobre el bien y hago un mejor uso del mismo.

	Uso del Vehículo		
	Privado (1)	No Privado (0)	Total
Nº Observaciones	**3,851**	**3,632**	**7,483**
Turismos (A)	*3,518*	*324*	*3,842*
No Turismos (G,M,O)	*333*	*3,308*	*3,641*

Nº Siniestros Anuales	**611**	**408**	**1,020**
Turismos (A)	*560*	*78*	*638*
No Turismos (G,M,O)	*52*	*330*	*382*
Media	**0.16**	**0.11**	**0.14**
Turismos (A)	*0.16*	*0.24*	*0.17*
No Turismos (G,M,O)	*0.16*	*0.10*	*0.10*

Tabla 9. Tabla del factor de riesgo uso de vehículo y tipo. Fuente: Elaboración propia.

3.4.4. Bonificación

El historial del cliente es un factor de riesgo cuantitativo que podrá tomar 6 valores: 0, 10, 20, 30, 40 y 50. En este caso representará la bonificación obtenida por el tomador en caso de presentar un buen historial de conducción.

La distribución de las observaciones para este factor de riesgo es la siguiente:

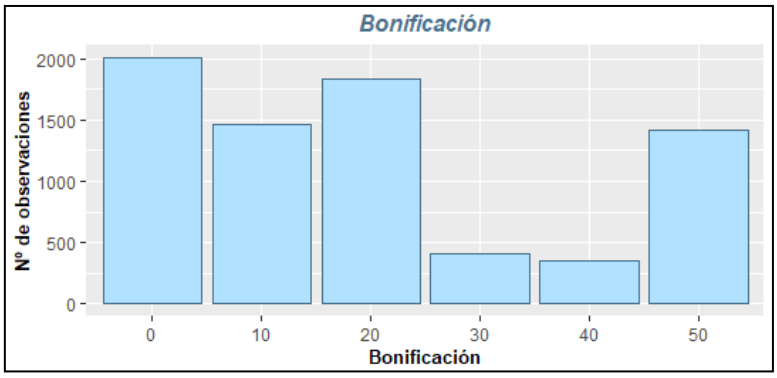

Gráfico 4. Factor de riesgo: Bonificación. Fuente: Elaboración propia.

La compañía aseguradora ha decidido agrupar los riesgos aplicando bonificaciones de diez en diez. No obstante, esta agrupación podría ser más o menos granular en función del historial. En este caso, para ofrecer consistencia en los resultados se agruparán las bonificaciones 30 y 40 con las 20 y 50 respectivamente. De esta forma, tal y como se observa en el Gráfico 5 la distribución de pólizas resulta más equitativa en lo que a volumen del portafolio se refiere:

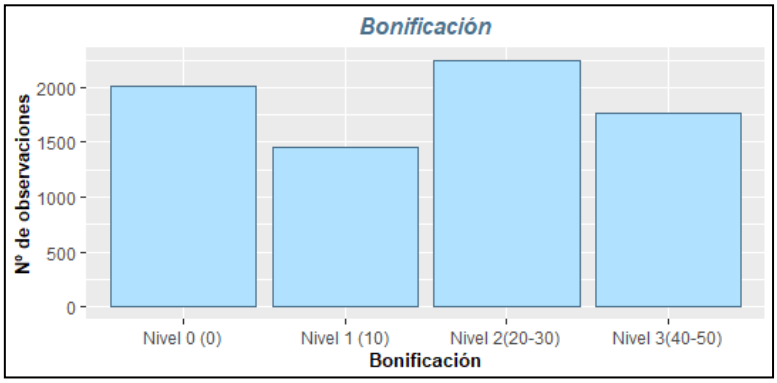

Gráfico 5. Factor de riesgo: Bonificación por niveles. Fuente: Elaboración propia.

En la Tabla 10 se puede observar como a medida que aumenta el nivel de bonificación la media de siniestros es inferior. Este resultado va en linea de lo que cabría esperar pues un cliente cuyo historial de conducción ya presenta frecuencias elevadas es probable que continúe declarando siniestros.

	Bonificación				
	Nivel 0 (0)	Nivel 1 (10)	Nivel 2 (20-30)	Nivel 3 (40-50)	Total
Nº Observaciones	2,010	1,458	2,247	1,768	**7,483**
% Obs.	26.86%	19.48%	30.03%	23.63%	**100.00%**

Nº Siniestros Anuales	**367**	**193**	**260**	**199**	**1,020**
Media	0.18	0.13	0.12	0.11	0.14
Mediana	0.00	0.00	0.00	0.00	0.00
Desviación estándar	0.77	0.81	0.65	0.61	0.71
Varianza	0.59	0.65	0.43	0.37	0.50
Asimetría	7.54	17.50	10.40	10.40	11.64
Kurtosi	81.47	444.90	150.85	160.87	229.51
Min	0.00	0.00	0.00	0.00	0.00
Max	11.78	22.83	11.78	11.78	22.83
Cuantil .25	0.00	0.00	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00	0.00	0.00
Cuantil .95	1.39	1.09	1.00	1.02	1.17

Tabla 10. Tabla del factor de riesgo bonificación. Fuente: Elaboración propia.

3.4.5. Edad del Tomador (AgeCat)

La edad del tomador es variable cuantitativa agrupada y categorizada en 7 niveles de respuesta: hasta 21 años, de 22 a 25, de 26 a 35, de 36 a 45, de 46 a 55, de 56 a 65 y más de 65 años.

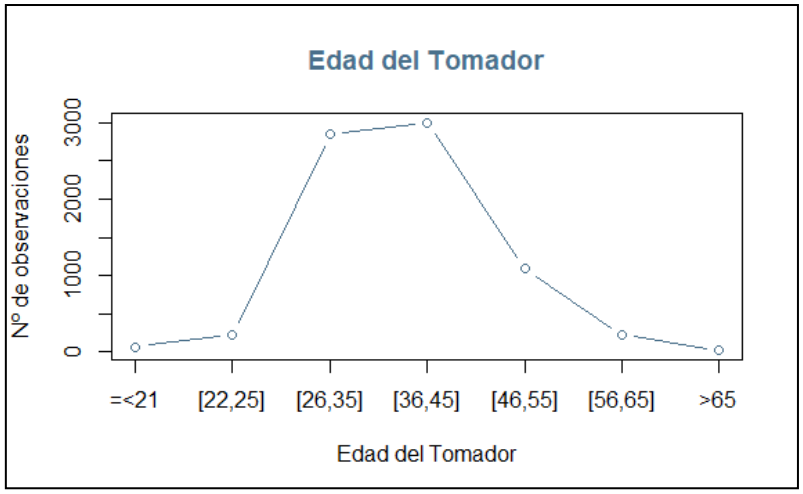

Gráfico 6. Factor de riesgo: Edad del Tomador. Fuente: Elaboración propia.

No obstante, tal y como se puede observar en el Gráfico 6 los rangos de edad inferiores a 25 años y superiores a 56 años presentan un peso muy reducido en la muestra. Por dicho motivo se realiza la siguiente reagrupación: hasta 25 años, de 26 a 35, de 36 a 45, de 46 a 55 y más de 56 años, resultando una nueva distribución del portafolio tal y como se puede visualizar en el Gráfico 7.

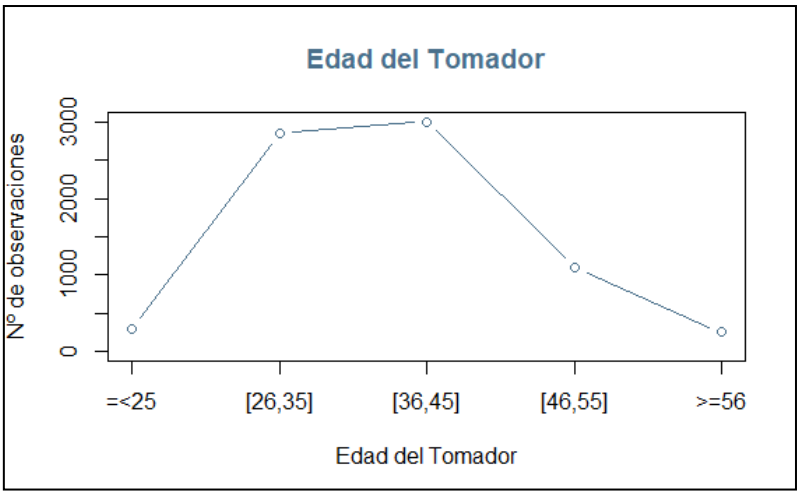

Gráfico 7. Factor de riesgo: Edad del Tomador Agrupado. Fuente: Elaboración propia.

En este sentido la Tabla 11 presenta una media de siniestros anuales significativamente superior inferiores a 25 años y superiores a 56 años. En lo que se refiere al primer grupo esto es algo que cabría esperar debido a la falta de experiencia. Esto impacta en cierta medida también en el grupo de 26 a 35 años donde también podría haber perfiles con reducida experiencia. Por su parte en el grupo de más de 56 años de edad el aumento de frecuencia también sería esperable debido al aumento de pérdida de facultades en la conducción a medida que se incrementa la edad.

	Edad del Tomador					
	=<25	26-35	36-45	46-55	>56	Total
Nº Observaciones	295	2,846	3,002	1,092	248	**7,483**
% Obs.	3.94%	38.03%	40.12%	14.59%	3.31%	**100.00%**

Nº Siniestros Anuales	**58**	**473**	**346**	**94**	**49**	**1,020**
Media	0.20	0.17	0.12	0.09	0.20	0.14
Mediana	0.00	0.00	0.00	0.00	0.00	0.00
Desviación estándar	0.73	0.84	0.66	0.41	0.66	0.71
Varianza	0.53	0.70	0.43	0.17	0.44	0.50
Asimetría	4.63	12.22	10.75	5.40	3.96	11.64
Kurtosi	24.61	241.01	158.20	31.66	17.05	229.51
Min	0.00	0.00	0.00	0.00	0.00	0.00
Max	5.62	22.83	11.78	3.97	4.45	22.83
Cuantil .25	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .95	1.64	1.20	1.00	0.00	1.33	1.17

Tabla 11. Tabla del factor de riesgo Edad. Fuente: Elaboración propia.

3.4.6. Antigüedad de Vehículo (VAgeCat)

La antigüedad del vehículo es una variable cuantitativa agrupada y categorizada en 7 niveles de respuesta: 0 años o nuevos, 1 año, 2 años, de 3 a 5 años, de 6 a 10 años, de 11 a 15 años o más de 16 años.

En el Gráfico 8 se puede observar que hay un volumen significativo de vehículos nuevos. Para el resto de las antigüedades el peso es similar, salvo para el último grupo, que por el momento no se agrupará a la espera de ver los resultados que ofrezcan los modelos de frecuencia.

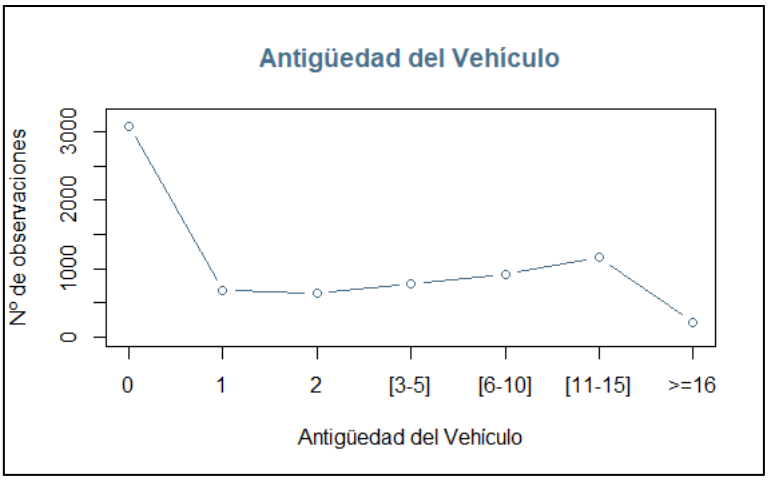

Gráfico 8. Factor de riesgo: Antigüedad del Vehículo. Fuente: Elaboración propia.

En la Tabla 12 se puede observar como a medida que aumenta la antigüedad disminuye la frecuencia. Cabría esperar que a mayor antigüedad hubiese una menor contratación de garantías y que por tanto el número de siniestros también fuese menor.

En este sentido, los vehículos hasta 5 años tienden a contratar garantias a todo riesgo. La inclusión de la garantía de daños propios haría aumentar la frecuencia de forma significativa. Adicionalmente, es bastante común que en vehículos de hasta 2 años de antigüedad la modalidad contratada sea a todo riesgo sin franquicia y el cuidado del vehículo será mayor. Esto generaría que a lo largo del segundo año aumente la frecuencia siniestral sabiendo que al año siguiente al incorporar una franquicia el coste para el tomador al declarar un siniestro será superior.

	Antigüedad del Vehículo							
	0	1	2	3-5	6-10	11-15	>16	Total
Nº Observaciones	3,079	681	646	771	924	1,164	218	**7,483**
% Obs.	41.15%	9.10%	8.63%	10.30%	12.35%	15.56%	2.91%	**100.00%**

Nº Siniestros Anuales	**484**	**119**	**154**	**120**	**89**	**49**	**5**	**1,020**
Media	0.16	0.17	0.24	0.16	0.10	0.04	0.02	0.14
Mediana	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Desviación estándar	0.78	0.73	1.14	0.69	0.46	0.28	0.16	0.71
Varianza	0.61	0.53	1.30	0.48	0.21	0.08	0.03	0.50
Asimetría	8.66	8.05	13.39	8.71	6.38	7.83	7.17	11.64
Kurtosi	101.22	101.34	243.34	115.04	51.48	71.99	49.93	229.51
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	11.78	11.78	22.83	11.78	5.99	3.97	1.33	22.83
Cuantil .25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cuantil .95	1.24	1.37	1.39	1.22	1.00	0.00	0.00	1.17

Tabla 12. Tabla del factor de riesgo Antigüedad del Vehículo. Fuente: Elaboración propia.

3.5. Análisis de sobredispersión

Para interpretar qué modelo ajustaría de forma más adecuada los datos se realizará en primer lugar un análisis de dispersión para entender si existe sobredispersión en los datos. Tal y como se ha comentado a lo largo del presente trabajo la regresión de Poisson supone

una misma media y varianza de la frecuencia siniestral por tanto supondrá equidispersión en los datos. Por tanto, en caso de que se observe sobredispersión importante en los datos podríamos concluir que el modelo de Poisson no sería aplicable para nuestros datos.

Para realizar un primer análisis se utilizará la función "dispersiontest" del paquete de R "AER" disponible para el modelo de Poisson de Cameron y Trivedi (1990).

Este test valida si existe dispersión a través de una esperanza E(Y)=μ y una varianza V(Y)=μ+alpha. Adicionalmente gracias a la variable alpha se puede conocer si los datos tienden a tener sobredispersión con alpha > 0 o infradispersión con alpha < 0. Por tanto, en caso de alpha>0 la varianza será mayor a la esperanza siniestral y los datos reflejarán sobredispersión.

Esta función parte de la Hipotesis Nula (H0) de equidispersión. En este caso el p-valor superior a 0.05 por tanto no se rechaza la hipótesis nula de datos equidispersos.

```
        Overdispersion test

data:  Pois_reg
z = 0.5036, p-value = 0.3073
alternative hypothesis: true alpha is greater than 0
sample estimates:
      alpha
0.007140932
```

No obstante, se puede observar un valor de alpha positivo por lo que habrá una tendencia a la sobredispersión.

Adicionalmente a este análisis podremos identificar sobredispersión relacionando el estadístico chi-cuadrado de Pearson (χ^2) o la función de desviación (D) con los grados de libertad del modelo (gl) (Vives, 2002).

Identificados ambos estadísticos podrá identificarse sobredispersión en caso de:

$$\frac{\chi^2}{gl} \ \ o\ también\ \frac{D}{gl}$$

Sean > 1.

El estadístico chi-cuadrado de Pearson (χ^2) compara el modelo teórico con el modelo observado, por lo que a menor valor mejor ajuste. Siendo su formulación:

$$\chi^2 = \sum_{1=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

siendo i el número de observaciones.

Por su parte la función o estadístico de desviación:

$$D(y;\hat{\mu}) = 2\{l(y;y) - l(\hat{\mu};y)\}$$

Que permitirá comprar el modelo aplicado con la función de verosimilitud de un modelo donde el número de parámetros estimado es igual al número de observaciones (saturado). Por tanto, cuando menor sea el valor de la desviación nos indicará que para un número reducido de parámetros se obtendrá un ajuste parecido al realizado cuando se ajuste en un modelo saturado con muchos parámetros.

Los grados de libertad se calcularán con el número de observaciones menos el número de parámetros estimados. Cada parámetro para del modelo reducirá en uno los grados de libertad.

En los modelos de Poisson y Binomial negativa aplicados se ha obtenido:

Poisson: Residual deviance: 2544.9 on 7464 degrees of freedom

Binomial Negativa: Residual deviance: 2457.5 on 7464 degrees of freedom

Observando los datos obtenidos en ambos modelos podemos observar de nuevo que no existe una dispersión muy significativa de los datos al no obtener un cociente superior a 1 que marcaría una sobredispersión significativa.

Esto hace que el modelo de Poisson sea válido para nuestros datos lo que nos llevará a comparar las bondades de ajuste y criterios de información de los modelos.

4. Especificación, análisis e interpretación de los modelos

En nuestra base de datos dispondremos de 6 factores de riesgo que serán variables categóricas ya que las edades y antigüedades se encuentran agrupadas.

Las variables Género y Uso estarán formadas por 2 niveles cada una, las variables Grupo Tipo de Vehículo y Nivel de Bonificación estarán formadas por 4 variables cada una de ellas, la variable AgeCat (edad) estará formada por 5 niveles y finalmente la variable VAgeCat estará formada por 7 niveles. En este caso habrá por tanto 2x2x4x4x5x7=2.240 tipos de riesgo diferentes en función de la combinación de factores de riesgo. Esto hará necesario calcular la frecuencia para las 2.240 posibles combinaciones de los factores de riesgo a través de la exposición de las pólizas y el número de siniestros.

En la base de datos disponemos de la exposición (t_i) a través del campo Exp_Weight y del número de siniestros (y_i) a través de la variable Clm_Count.

En este caso, se identifica una relatividad para cada factor de tarificación igual a uno y que será el valor de base, tal y como se puede observar en la Tabla 13.

Con el valor base para cada factor de tarificación identificado como f_0 el resto de las relatividades se determinarán de una única forma en base a la relatividad de tarifa base. De esta forma f_0 representará el valor esperado para la tarifa base (Frees, 2020).

A la hora de formular el modelo de regresión cada opción de los distintos factores de tarificación llevará asociada una variable indicadora x_i. Teniendo en cuenta las distintas

variables explicativas en nuestra cartera de pólizas se podremos obtener incorporando la exposición m:

$$\log \mu i = x_i'\beta + \log t_i$$
$$= \log t_i + \beta_0 + \beta_1 I(Genero_i = M) + \beta_2 I(Uso_i = No\ Privado)$$
$$+ \sum_{t=3}^{5} \beta_t I(Vtipoveh_i = t-1) + \sum_{t=6}^{8} \beta_t I(Vncd_i = t-4)$$
$$+ \sum_{t=9}^{12} \beta_t I(Vedad_i = t-7) + \sum_{t=13}^{18} \beta_t I(Vantig._i = t-11)$$

Variable	*Opciones*	Nivel	f_0
Genero	**Mujer (F)**	**1**	**Base**
	Hombre (M)	2	
Uso	**Privado (1)**	**1**	**Base**
	No Privado (0)	2	
Grupo Tipo Vehículo	**Turism**	**1**	**Base**
	Furgonetas (G)	2	
	Motocicletas	3	
	Otros (O)	4	
Nivel Bonificación	**Nivel 0 (0)**	**1**	**Base**
	Nivel 1 (10)	2	
	Nivel 2 (20-30)	3	
	Nivel 3 (40-50)	4	
AgeCat	**36-45**	**1**	**Base**
	≤25	2	
	26-35	3	
	46-55	4	
	≥ 56	5	
VAgeCat	**0**	**1**	**Base**
	1	2	
	2	3	
	3-5	4	
	6-10	5	
	11-15	6	
	≥ 16	7	

Tabla 13. Tabla asignación opciones base. Fuente: Elaboración propia.

4.1. Regresión de Poisson

Tal y como se menciona en el apartado 1.2.2. podremos disponer del parámetro de tasa media de Poisson calculado como $\lambda_i = \frac{y_i}{t_i}$.

En la regresión de Poisson se calculará la frecuencia haciendo uso de la tasa o media λ y exposición t se podrá conocer el número de siniestros esperado en función de la exposición como $t\lambda$. En este caso, t es una variable conocida y estaremos interesado en modelizar λ en función de los factores de riesgo.

La tasa Poisson se podrá modelar como:

$\lambda = f_0 x f_{1,genero} x f_{2,uso} x f_{3,tipoveh}\ x f_{4,Bonif}\ x f_{5,edad}\ x f_{6,antig.}$

Pudiéndose reescribir como logaritmo:

$\log \lambda = logf_0 + logf_{1,genero} + logf_{2,uso} + logf_{3,tipoveh} + logf_{4,Bonif} + logf_{5,edad} + logf_{6,antig.}$

Se lanza el modelo a través de la fórmula de R Studio:

Modelo de Regresión de Poisson					
Variable	***Opciones***	**Estimate**	**Std. Error**	**Pr(>|z|)**	
Intercepto		-2.026	0.164	0.000	
Genero	**Mujer (F)**				
	Hombre (M)	0.291	0.119	0.014	*
Uso	**Privado (1)**				
	No Privado (0)	-0.325	0.181	0.073	.
Grupo Tipo Vehículo	**Turism Automóvil (A)**				
	Furgonetas (G)	0.422	0.237	0.075	.
	Motocicletas (M)	-0.988	0.564	0.080	.
	Otros (O)	0.462	0.262	0.077	.
Nivel Bonificación	**Nivel 0 (0)**				
	Nivel 1 (10)	-0.323	0.127	0.011	*
	Nivel 2 (20-30)	-0.349	0.119	0.003	**
	Nivel 3 (40-50)	-0.571	0.134	0.000	***
AgeCat	≤25	0.825	0.200	0.000	***
	26-35	0.373	0.110	0.001	***
	36-45				
	46-55	-0.448	0.166	0.007	**
	≥ 56	0.505	0.216	0.019	*
VAgeCat	**0**				
	1	0.178	0.148	0.231	
	2	0.447	0.141	0.001	**
	3-5	0.204	0.204	0.317	
	6-10	-0.316	0.215	0.142	
	11-15	-1.113	0.249	0.000	***
	≥ 16	-1.585	0.524	0.002	**
AIC: 3578					
Number of Fisher Scoring iterations: 6					

Tabla 14. Tabla Modelo Poisson. Fuente: Elaboración propia.

Este modelo parte de:

- **Hipótesis Nula (H0):** el factor de riesgo no presenta impacto en la frecuencia siniestral al analizar su influencia manteniendo el resto de las variables invariables.

- **Hipótesis Alternativa (H1):** rechaza la hipótesis nula de que el valor estimado del factor de riesgo pueda ser igual a 0.

Para interpretar sus resultados emplearemos el p-valor. Si el p-valor es menor que o igual al nivel de significancia que se determine se podrá rechazar la hipótesis nula y podremos concluir que el factor de riesgo presenta un impacto significativo en la frecuencia

siniestral. Una vez determinado que sí es significativo si el valor del estimador presenta valores positivos la frecuencia se espera que sea superior la del valor de referencia, en caso de ser negativa será inferior.

Al aplicar el modelo se obtiene la estimación, el error estándar y el p-valor, tal y como se observa en la Tabla 14.

A través de los resultados obtenidos se puede observar que:

- En género masculino presenta un valor de estimación positivo (0.291) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base. Adicionalmente el p-valor muestra un valor de 0.014 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0, por tanto, con un nivel de confianza del 95% se puede descartar que el valor será 0. Por tanto, el género sí será significativo en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En uso no privado presenta un valor de estimación negativo (-0.325) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base de uso privado. El p-valor muestra un valor de 0.073 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza 90%. Podemos concluir que uso sí será significativo en la estimación de la siniestralidad media a un nivel de confianza del 90%, no así al 95%

- En relación con el grupo de tipo de vehículo podemos observar:

    - Las furgonetas (G) presenta un valor de estimación positivo (0.422) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.075 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza del 90%.

    - Las motocicletas (G) presenta un valor de estimación negativo (-0.988) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.080 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 un nivel de confianza del 90%.

    - El resto de los tipos de vehículo (O) presenta un valor de estimación positivo (0.462) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.077 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza del 90%.

      Podemos concluir por tanto que el tipo de vehículo sí será significativo en la estimación de la siniestralidad media a un nivel de confianza del 90% de confianza, no así al 95%

- Para los niveles de Bonificación:

    - El resultado de bonificación de Nivel 1 (10) presenta un valor de estimación negativo (-0.323) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.011 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - El resultado de bonificación de Nivel 2 (20-30) presenta un valor de estimación negativo (-0.349) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.003 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - El resultado de bonificación de Nivel 3 (40-50) presenta un valor de estimación negativo (-0.571) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

      Podemos concluir por tanto que la bonificación sí será significativa en la estimación de la siniestralidad media a un nivel de confianza del 95%.

- Para los siguientes rangos de edad, partiendo de un rango de 36 a 45 años como base por su elevado peso, podemos concluir:

    - Para edades < 25 años se observa un valor de estimación positivo (0.825) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - Para edades de 26 a 35 años se observa un valor de estimación positivo (0.373) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.001 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - Para edades de 45 a 55 años se observa un valor de estimación negativo (-0.448) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.007 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - Para edades a partir de 56 años se observa un valor de estimación positivo (0.505) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.019 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Para los siguientes rangos de antigüedad de vehículo, partiendo de un vehículo nuevo (0 años) como base por su elevado peso en la muestra, podemos concluir:

    - Con la antigüedad de 1 año se puede observar un valor de estimación positivo (0.178) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.231 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 1 año no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

    - Con la antigüedad de 2 años se observa un valor de estimación positivo (0.447) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.001 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - Con la antigüedad de 3 a 5 años se puede observar un valor de estimación positivo (0.204) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.317 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 3 a 5 años no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

    - Con la antigüedad de 6 a 10 años se puede observar un valor de estimación negativo (-0.316) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.142 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 6 a 10 años no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

    - Con la antigüedad de 11 a 15 años se puede observar un valor de estimación negativo (-1.113) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    - Con la antigüedad a partir de 16 años se puede observar un valor de estimación negativo (-1.585) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.002 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

Por otro lado, el modelo también nos devolverá las relatividades que serán los resultados del ajuste a aplicar.

Variable	*Opciones*	Relatividad
Genero	**Mujer (F)**	**1**
	Hombre (M)	1.338
Uso	**Privado (1)**	**1**
	No Privado (0)	0.722
Grupo Tipo Vehículo	**Turism Automóvil (A)**	**1**
	Furgonetas (G)	1.525
	Motocicletas (M)	0.372
	Otros (O)	1.588
Nivel Bonificación	**Nivel 0 (0)**	**1**
	Nivel 1 (10)	0.724
	Nivel 2 (20-30)	0.705
	Nivel 3 (40-50)	0.565
AgeCat	**36-45**	**1**
	≤25	2.281
	26-35	1.452
	46-55	0.639
	≥ 56	1.657
VAgeCat	**0**	**1**
	1	1.194
	2	1.564
	3-5	1.226
	6-10	0.729
	11-15	0.328
	≥ 16	0.205

Tabla 15. Tabla Modelo Poisson Relatividades. Fuente: Elaboración propia.

Tal y como refleja la Tabla 15 podemos observar:

- La frecuencia siniestral es un 33% superior en hombres que en mujeres manteniendo fijos el resto de los factores de tarificación. Esto ya se podía observar en el análisis descriptivo donde se reflejaba una mayor media siniestral en la Tabla 5.

- La frecuencia siniestral en caso de vehículos de uso no privado es inferior a la frecuencia de uso privado. En este caso el uso no privado es un 72% de la frecuencia que representa el uso privado. En la Tabla 9 se puede observar cómo este valor se acentúa en furgonetas donde la frecuencia en caso de uso no privado se reduce. Esto podría deberse fundamentalmente a un mayor cuidado en propiedades de empresa o bien a posibles sanciones económicas implícitas en el uso del vehículo por la propia empresa.

- En cuanto al tipo de vehículo podemos observar cómo las Furgonetas y Resto de Vehículos presentan una frecuencia siniestral un 52,5% y 58,8% respectivamente, superior a la frecuencia observada en turismos. Esto viene explicado por un mayor uso del vehículo circulando entorno a las 8 horas diarias.
  En cuanto a las motocicletas estas representan un muy buen resultado de frecuencia siniestral obteniendo un 37% de la frecuencia obtenida en turismos. Esto podría deberse a una selección de riesgos más acotada por parte de la compañía.

- En cuanto a la bonificación la frecuencia siniestral disminuye a medida que aumenta la bonificación. En este caso, además el nivel 3 (40-50) que representa la mayor

bonificación presenta una frecuencia del 56,5% de la frecuencia obtenida en la opción básica sin bonificación.

- En relación con la edad se puede observar cómo los perfiles de edad inferior presentan incrementos de frecuencia. Este aumento vendría explicado por la falta de experiencia de los conductores del vehículo.
  Por otro lado, la edad con un rango superior de 46 a 55 años presenta una mejora de frecuencias. Sin embargo, a partir de los 56 años se empezarían a observar las consecuencias de la pérdida de reflejos en la conducción aumentando a partir de dicha edad un 65,7% la frecuencia siniestral respecto a la frecuencia de referencia de 36 a 45 años).

- En cuanto a la antigüedad del vehículo se puede observar que entre los años 1 y 5 se observaría un aumento de la frecuencia siniestral. Este aumento vendría explicado por la contratación de un mayor número de garantías, especialmente la garantía de daños propios ya que en este caso se busca mantener el vehículo totalmente cubierto.
  Adicionalmente, se puede observar un repunte de la frecuencia en el año 2. Esto vendría explicado por la declaración de siniestros a lo largo del segundo año de antigüedad pues es frecuente la contratación de franquicias a partir del tercer año.
  En contra posición a partir de los 6 años la frecuencia disminuiría de forma significativa, especialmente en vehículos a partir de 16 años donde la frecuencia observada para dicho perfil manteniendo el resto de los factores invariables es un 20,5% de lo observado para el valor de referencia de 0 años.

4.2. Regresión Binomial Negativa

El modelo parte de:

- **Hipótesis Nula (H0):** el factor de riesgo no presenta impacto en la frecuencia siniestral al analizar su influencia manteniendo el resto de las variables invariables.

- **Hipótesis Alternativa (H1):** rechaza la hipótesis nula de que el valor estimado del factor de riesgo pueda ser igual a 0.

Como sucedía con el modelo de regresión de Poisson para interpretar sus resultados emplearemos el p-valor. Si el p-valor es menor que o igual al nivel de significancia que se determine se podrá rechazar la hipótesis nula y podremos concluir que el factor de riesgo presenta un impacto significativo en la frecuencia siniestral. Unas cifras positivas en los estimadores, una vez determinado que el factor de riesgo es significativo, harán que la frecuencia esperable sea superior la del valor de referencia, en caso de ser negativa será inferior.

Al aplicar el modelo se obtiene la estimación, el error estándar y el p-valor, tal y como se observa en la Tabla 16.

Modelo de Regresión de Binomial Negativa					
Variable	***Opciones***	**Estimate**	**Std. Error**	**Pr(>|z|)**	
Intercepto		-2.016	0.165	0.000	
Género	**Mujer (F)**				
	Hombre (M)	0.285	0.120	0.017	*
Uso	**Privado (1)**				
	No Privado (0)	-0.320	0.183	0.080	.
Grupo Tipo Vehículo	**Turism Automóvil (A)**				
	Furgonetas (G)	0.394	0.240	0.100	
	Motocicletas (M)	-1.002	0.567	0.077	.
	Otros (O)	0.449	0.264	0.089	.
Nivel Bonificación	**Nivel 0 (0)**				
	Nivel 1 (10)	-0.325	0.129	0.012	*
	Nivel 2 (20-30)	-0.347	0.120	0.004	**
	Nivel 3 (40-50)	-0.576	0.135	0.000	***
AgeCat	≤25	0.813	0.205	0.000	***
	26-35	0.369	0.111	0.001	***
	36-45				
	46-55	-0.437	0.167	0.009	**
	≥ 56	0.517	0.218	0.018	*
VAgeCat	**0**				
	1	0.177	0.150	0.239	
	2	0.447	0.142	0.002	**
	3-5	0.211	0.207	0.308	
	6-10	-0.303	0.218	0.165	
	11-15	-1.099	0.252	0.000	***
	≥ 16	-1.571	0.525	0.003	**
AIC: 3579.1					
Number of Fisher Scoring iterations: 1					

Tabla 16. Modelo Binomial Negativa. Fuente: Elaboración propia.

A través de los resultados obtenidos se puede observar que:

- En género masculino presenta un valor de estimación positivo (0.285) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base. Adicionalmente el p-valor muestra un valor de 0.017 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0, por tanto, con un nivel de confianza del 95% se puede descartar que el valor será 0. El género sí será significativo en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En uso no privado presenta un valor de estimación negativo (-0.320) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base de uso privado. El p-valor muestra un valor de 0.080 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza del 90%. Podemos concluir que el uso sí será significativo en la estimación de la siniestralidad media a un nivel de confianza del 90%, no así al 95%

- En relación con el grupo de tipo de vehículo podemos observar:

    - Las furgonetas (G) presenta un valor de estimación positivo (0.394) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.1002 por lo que no podemos

rechazar la hipótesis de que el valor pueda ser 0 ni a un nivel de confianza del 95% ni del 90%.

- Las motocicletas (G) presenta un valor de estimación negativo (-1.002) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.077 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza 90%.

- El resto de los tipos de vehículo (O) presenta un valor de estimación positivo (0.449) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base de turismos. El p-valor muestra un valor de 0.089 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%. Si que se podrá rechazar la hipótesis de que el valor pueda ser 0 a un nivel de confianza 90%.

  Podemos concluir por tanto que el tipo de vehículo motocicleta y otro tipo de vehículo sí será significativo en la estimación de la siniestralidad media a un nivel de confianza 90% de confianza, no así al 95%. Sin embargo, se podrá considerar el mismo comportamiento para turismos que para furgonetas.

- Para los niveles de Bonificación:

  - El resultado de bonificación de Nivel 1 (10) presenta un valor de estimación negativo (-0.325) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.012 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

  - El resultado de bonificación de Nivel 2 (20-30) presenta un valor de estimación negativo (-0.347) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.004 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

  - El resultado de bonificación de Nivel 3 (40-50) presenta un valor de estimación negativo (-0.576) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

    Podemos concluir por tanto que la bonificación sí será significativa en la estimación de la siniestralidad media a un nivel de confianza del 95%.

- Para los siguientes rangos de edad, partiendo de un rango de 36 a 45 años como base por su elevado peso, podemos concluir:
  - Para edades < 25 años se observa un valor de estimación positivo (0.813) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos

rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Para edades de 26 a 35 años se observa un valor de estimación positivo (0.369) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.001 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Para edades de 45 a 55 años se observa un valor de estimación negativo (-0.437) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.009 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Para edades a partir de 56 años se observa un valor de estimación positivo (0.517) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.018 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Para los siguientes rangos de antigüedad de vehículo, partiendo de un vehículo nuevo (0 años) como base por su elevado peso en la muestra, podemos concluir:

  - Con la antigüedad de 1 año se puede observar un valor de estimación positivo (0.177) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.239 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 1 año no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

  - Con la antigüedad de 2 años se observa un valor de estimación positivo (0.447) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.002 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

  - Con la antigüedad de 3 a 5 años se puede observar un valor de estimación positivo (0.211) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.308 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 3 a 5 años no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

  - Con la antigüedad de 6 a 10 años se puede observar un valor de estimación negativo (-0.303) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. No obstante, el p-valor muestra un valor de 0.165 por lo que no podemos rechazar la hipótesis de que el valor pueda ser 0. Por tanto, la antigüedad de 6 a 10 años no será significativa en la estimación de la siniestralidad media respecto al valor de base de 0 años.

- Con la antigüedad de 11 a 15 años se puede observar un valor de estimación negativo (-1.099) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.000 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

- Con la antigüedad a partir de 16 años se puede observar un valor de estimación negativo (-1.571) y, por tanto, el volumen de siniestros esperado será inferior que el obtenido en la opción base sin bonificación. El p-valor muestra un valor de 0.003 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0 con un nivel de confianza del 95%.

Por otro lado, el modelo también nos devolverá las relatividades que serán los resultados del ajuste a aplicar.

Variable	***Opciones***	**Relatividad**
Genero	**Mujer (F)**	**1**
	Hombre (M)	1.330
Uso	**Privado (1)**	**1**
	No Privado (0)	0.726
Grupo Tipo Vehículo	**Turism Automóvil (A)**	**1**
	Furgonetas (G)	1.483
	Motocicletas (M)	0.367
	Otros (O)	1.567
Nivel Bonificación	**Nivel 0 (0)**	**1**
	Nivel 1 (10)	0.723
	Nivel 2 (20-30)	0.707
	Nivel 3 (40-50)	0.562
AgeCat	≤25	2.254
	26-35	1.447
	36-45	**1**
	46-55	0.646
	≥ 56	1.677
VAgeCat	**0**	**1**
	1	1.193
	2	1.563
	3-5	1.234
	6-10	0.739
	11-15	0.333
	≥ 16	0.208

Tabla 17. Tabla Modelo Binomial Negativa Relatividades. Fuente: Elaboración propia.

Tal y como refleja la Tabla 17 podemos observar:

- La frecuencia siniestral es un 33% superior en hombres que en mujeres manteniendo fijos el resto de los factores de tarificación. Este porcentaje reflejará el mismo resultado que la regresión del modelo de Poisson.

- La frecuencia siniestral en caso de vehículos de uso no privado es inferior a la frecuencia de uso privado. En este caso el uso no privado es un 72% de la frecuencia

que representa el uso privado. Este porcentaje también se reflejaba en la regresión del modelo de Poisson.

- En cuanto al tipo de vehículo se observa que Furgonetas y Resto de Vehículos presentan una frecuencia siniestral un 48,3% y 56,7% respectivamente, superior a la frecuencia observada en turismos. No obstante, en las furgonetas los resultados no parecían ser significativos. En cuanto a las motocicletas estas representan una frecuencia siniestral obteniendo un 37% de la frecuencia obtenida en turismos.

- En cuanto a la bonificación la frecuencia siniestral disminuye a medida que aumenta la bonificación como ya sucedía en la regresión del modelo de Poisson.

- En relación con la edad se puede observar cómo los perfiles de 46 a 55 años presentan una mejora de frecuencias. El resto de las edades presenta un deterioro de frecuencia siniestral respecto a la edad de referencia.

- En cuanto a la antigüedad del vehículo se puede observar que a partir de los 6 años la frecuencia disminuiría de forma significativa algo que ya sucedía en el modelo de Poisson, especialmente en vehículos a partir de 16 años donde la frecuencia es un 20,8% de lo observado para el valor de referencia de 0 años.

4.3. Regresión Poisson cero inflado

Tal y como se observaba en el apartado Tabla 4 del apartado 3.3. en la distribución de pólizas se puede observar un volumen importante de clientes con 0 siniestros. Esto es debido, como ya se comentaba, a que existe volumen importante de ceros debido a que habrá una parte de conductores que no hayan utilizado el vehículo y, que por tanto su frecuencia siniestral será cero, y la segunda de ellas vendrá motivada por una parte de conductores del portafolio con buenos hábitos de conducción.

La regresión de Poisson cero inflado está formada por un componente cero inflado y un componente de conteo. A través del componente binario o componente de cero inflado podremos saber en caso de obtener un signo positivo en el estimador del valor de referencia aumentará la probabilidad de una frecuencia siniestral de cero generada por la primera fuente de cero (Yoshika, 2014).

Este componente parte de la **Hipótesis Nula (H0)** de tener 0 siniestros. Por tanto, si el resultado al aplicar el modelo es significativo aumentarán las probabilidades de encontrar una frecuencia siniestral igual a 0.

Por otra parte, en el componente de conteos al seguir una distribución una Poisson será interpretada de la misma forma que la regresión de Poisson donde la **Hipótesis Nula (H0)** es que el factor de riesgo podrá tener un valor estimado de 0.

Si aplicamos el modelo de regresión de Poisson cero inflado obtendremos:

Modelo de Regresión de Poisson Cero Inflado					
Componente de Cero inflado (binomial con logit link)					
Variable	***Opciones***	**Estimate**	**Std. Error**	**Pr(>|z|)**	
Intercepto		-40.992	121.332	0.735	
Genero	**Mujer (F)**				
	Hombre (M)	13.901	60.624	0.819	
Uso	**Privado (1)**				
	No Privado (0)	4.878	0.972	0.000	***
Grupo Tipo Vehículo	**Turism Automóvil (A)**				
	Furgonetas (G)	13.746	60.580	0.821	
	Motocicletas (M)	19.397	87.575	0.825	
	Otros (O)	11.888	60.573	0.844	
Nivel Bonificación	**Nivel 0 (0)**				
	Nivel 1 (10)	1.663	0.622	0.007	**
	Nivel 2 (20-30)	-0.007	0.516	0.989	
	Nivel 3 (40-50)	21.690	60.627	0.721	
AgeCat	≤25	-5.776	1.149	0.000	***
	26-35	-4.785	0.906	0.000	***
	36-45				
	46-55	-37.762	169.258	0.823	
	≥ 56	-32.711	148.392	0.826	
VAgeCat	**0**				
	1	2.587	3.158	0.413	
	2	6.219	2.386	0.009	**
	3-5	8.766	3.240	0.007	**
	6-10	8.587	3.246	0.008	**
	11-15	29.836	60.886	0.624	
	≥ 16	-5.717	273.604	0.983	
Componente de Conteo (Poisson con log link)					
Variable	***Opciones***	**Estimate**	**Std. Error**	**Pr(>|z|)**	
Intercepto		-2.084	0.166	0.000	***
Genero	**Mujer (F)**				
	Hombre (M)	0.480	0.130	0.000	***
Uso	**Privado (1)**				
	No Privado (0)	0.006	0.179	0.975	
Grupo Tipo Vehículo	**Turism Automóvil (A)**				
	Furgonetas (G)	0.311	0.233	0.182	
	Motocicletas (M)	-1.317	0.567	0.020	*
	Otros (O)	0.145	0.268	0.588	
Nivel Bonificación	**Nivel 0 (0)**				
	Nivel 1 (10)	-0.189	0.132	0.153	
	Nivel 2 (20-30)	-0.341	0.126	0.007	**
	Nivel 3 (40-50)	-0.425	0.141	0.003	**
AgeCat	≤25	0.502	0.208	0.016	
	26-35	0.140	0.115	0.221	
	36-45				
	46-55	-0.915	0.182	0.000	***
	≥ 56	0.088	0.221	0.690	
VAgeCat	**0**				
	1	0.217	0.150	0.149	
	2	0.541	0.147	0.000	***
	3-5	0.466	0.213	0.029	*
	6-10	0.002	0.240	0.994	
	11-15	1.724	0.316	0.000	***
	≥ 16	-1.542	0.524	0.003	**
Log-likelihood: -1718 on 38 Df					
Number of iterations in BFGS optimization: 230					

Tabla 18. Modelo Poisson cero inflado. Fuente: Elaboración propia.

A través de los resultados obtenidos se puede observar que:

- En género masculino será significativo en el componente de conteo y presenta un valor de estimación positivo (0,480) y, por tanto, el volumen de siniestros esperado será superior que el obtenido en la opción base. Adicionalmente el p-valor muestra un valor de 0 por lo que podemos rechazar la hipótesis de que el valor pueda ser 0, por tanto, con un nivel de confianza del 95% se puede descartar que el valor será 0. El género sí será significativo en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En uso no privado será significativo en el componente de cero inflado y dado que el valor de su estimador es positivo (4,878) indicará que si el vehículo es de uso no privado la probabilidad de encontrar 0 siniestros aumenta.

- El tipo de vehículo solo será significativo en caso de tratarse de furgonetas, contemplado en el componente del modelo de conteo. En este caso, con un valor del estimador negativo (-1,317) el volumen de siniestros esperado será inferior que el obtenido en la opción base. Adicionalmente el p-valor muestra un valor de 0,020 por lo que con un nivel de confianza del 95% se puede descartar que el valor será 0.
  El tipo de vehículo motocicleta sí será significativo en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En el nivel de Bonificación se puede observar que:

  - En el caso de disponer de una bonificación del 10% será significativo en el componente de cero inflado y dado que el valor de su estimador es positivo (1,663) indicará que si el cliente presenta bonificación del 10% la probabilidad de encontrar 0 siniestros aumenta.

  - Adicionalmente, el nivel de bonificación 2 (bonificación del 20% y 30%) y el nivel de bonificación 3 (40%-50%) serán significativos en el componente de conteo. Ambos presentarán un valor del estimador negativo (-0,341 y -0,425 respectivamente) por lo que el volumen de siniestros esperado será inferior que el obtenido en la opción base. Adicionalmente el p-valor muestra un valor de 0,007 y 0,003 respectivamente por lo que con un nivel de confianza del 95% se puede descartar que el valor será 0.

    Por todo ello, podemos determinar que los niveles de bonificación sí serán significativos en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En relación con la edad del tomador se puede concluir:

  - Las edades inferiores a 25 años y las comprendidas entre 25 y 35 años los factores de riesgo serán significativos en el componente de cero inflado y dado que el valor de su estimador es negativo (-5,776 y -4,785 respectivamente) indicará la probabilidad de encontrar 0 siniestros disminuye.

  - Adicionalmente, el rango de edad comprendido entre 46 y 55 años será significativos con un p-valor de 0 en el componente de conteo y presentará un

valor del estimador negativo (-0,915) por lo que el volumen de siniestros esperado será inferior que el obtenido en la opción base. Por tanto, con un nivel de confianza del 95% se puede descartar que el valor será 0.
Por todo ello, podemos determinar que las edades hasta 55 años sí serán significativas en la estimación de siniestralidad media a un nivel de confianza del 95%.

- En la antigüedad del vehículo se puede observar que:

  - Las antigüedades de 2 a 5 años serán significativas en el componente de cero inflado (6,219 y 8,766) y en el componente de conteo (0,541 y 0,466). Dado que todos los valores de sus estimadores son positivos esto indicará que si el vehículo presenta estas antigüedades la probabilidad de encontrar 0 siniestros aumenta, pero que a su vez cuando sí que se producen siniestros la frecuencia siniestral será superior a la observada cuando el vehículo es nuevo. Esto tiene sentido pues de 2 a 5 años los vehículos presentan un buen mantenimiento, pero a su vez suelen tener reducciones importantes de garantías fruto de la depreciación del vehículo incentivando la declaración antes de realizar la reducción.

  - Sí que se observarían mejoras claras de frecuencia para antigüedades de 6 a 10 donde aumenta la probabilidad de encontrar 0 siniestros, no siendo significativa en el componente de conteo.

  - Por su parte el riesgo para las antigüedades de 11 a 15 años resulta significativo en el componente de conteo con un valor del estimador positivo (1,724) por lo que el volumen de siniestros esperado será superior que el obtenido en la opción base.

  - Por otra parte, tendrán un buen comportamiento los vehículos de 16 años o más con una frecuencia siniestral inferior a la presentada por el valor de base tal y como indica el modelo de conteo.
    Descartamos que esto venga motivado por una falta de uso ya que no se observa significatividad en el modelo de cero inflado por tanto podemos asociar esta reducción de frecuencia a la reducción de garantías contratadas dado el poco valor que se le da al vehículo llegadas dichas antigüedades.

4.4. Regresión Binomial Negativa cero inflado

Si aplicamos el modelo de regresión de Binomial Negativa cero inflado vemos que el modelo no converge.

```
Warning message:
In value[[3L]](cond) :
  sistema es computacionalmente singular: número de condición recíproco = 1.23013e-34FALSE

Theta = 2135085.0301
```

Para nuestra base de datos ya se ha analizado la sobredispersión en el apartado 3.5. y podíamos observar cómo pese a existir cierta sobredispersión esta era muy reducida tal y como marcaba el valor de alpha. Adicionalmente, nos devuelve el valor de Theta cuyo resultado es muy elevado y será una clara señal que los datos no tienen sobredispersión

cuando se incluyes el cero inflado. Por tanto, no será necesario aplicar el modelo Binominal Negativo cero inflado y será mejor aplicar la Poisson cero inflado.

5. Ajuste y selección de los modelos

En primer lugar, es importante señalar que el modelo de regresión Binomial Negativa cero inflado es descartado debido a que no converge y adicionalmente la base de datos de nuestra cartera no presenta mucha dispersión. Por tanto, deberemos analizar y seleccionar el mejor modelo entre los restantes: regresión de Poisson, regresión Binomial Negativo y regresión de Poisson cero inflado.

Para iniciar este ejercicio se podrá analizar la bondad de ajuste a través del estadístico de desviación mencionado en el apartado 3.5. del presente trabajo:

$$D(y;\hat{\mu}) = 2\{l(y;y) - l(\hat{\mu};y)\}$$

Como ya se ha mencionado, cuanto menor sea su valor, como ya se ha comentado, se nos indicará que para un número reducido de parámetros se obtendrá un ajuste parecido al realizado cuando se ajuste en un modelo saturado con muchos parámetros.

Emplearemos este estadístico para comparar en primer lugar la regresión de Poisson y la regresión de Binominal Negativa obteniendo los resultados siguientes:

	Poisson	Binomial Negativa
Desviación	*2544.9*	*2457.5*

Tabla 19. Comparativa de la función de desviación. Fuente: Elaboración propia.

No podemos concluir que el modelo de regresión Binomial Negativa ofrezca un mejor ajuste que el modelo de Poisson. Esto puede deberse a la ligera sobredispersión existente en nuestros datos, lo cual da cabida a la regresión de Poisson cero inflado.

Por tanto, compararemos los tres modelos a través del criterio de información de Akaike (AIC) cuya formula es:

$$AIC = 2k - 2\ln(\hat{L})$$

Siendo k el número de parámetros y L es el máximo valor de la función de probabilidad para el modelo estimado.

El modelo obtendrá un valor de AIC reducido cuando el modelo se explique de la forma más adecuada con el menor número de parámetros posible. Por tanto, no sólo tendrá en cuenta la bondad de ajuste, es decir, lo bien que se pueda ajustar el modelo, sino que también penalizará la existencia de un mayor número de parámetros estimados.

También podemos comparar estos tres modelos a través del criterio de información Bayesiano (BIC), cuya formula es:

$$BIC = klog(n) - 2L$$

Siendo k el número de parámetros, L es el máximo valor de la función de probabilidad para el modelo estimado y n el número de observaciones.

De nuevo como sucedía en el criterio AIC, el modelo que presente un valor de BIC será el modelo seleccionado. El criterio BIC penalizará en mayor medida el uso de más regresores.

Aplicando ambos criterios obtendremos:

	Poisson	Binomial Negativa	Poisson Cero Inflado
AIC	*3578*	*3579*	*3513*
BIC	*3709*	*3717*	*3775*

Tabla 20. Comparativa del criterio AIC y BIC. Fuente: Elaboración propia.

En este caso, aplicando el criterio de información Akaike podemos determinar que el modelo que mejor se ajusta a los datos obteniendo un modelo de parámetros reducido será el modelo de Poisson cero inflado ya que es el modelo que simplifica el número de parámetros a través de Poisson y a su vez contempla la sobredispersión existente gracias a inflar ceros.
En cambio, por su parte el modelo BIC considera que la incorporación de mayor complejidad en el modelo construyendo el modelo de regresión de cero inflado no está justificado y será mejor escoger el modelo de regresión de Poisson.

Por tanto, se deberá escoger el uso de uno u otro modelo en función de qué objetivos se persiga y si se toma como primordial el uso de un menor número de regresores o en cambio se opta por un modelo con mayor volumen de regresores, pero mayor sofisticación en los resultados.

6. Conclusiones

A lo largo del presente trabajo se ha analizado y comprobado como los factores de riesgo presentan una incidencia importante sobre la frecuencia siniestral, empleando para dicho estudio una cartera de pólizas de automóviles de Singapur ofrecida por Frees (2010).

Esto ya parece reflejarse en un primer análisis descriptivo de los distintos factores de riesgo donde observa como la media siniestral, curtosis o asimetría, entre otros, varía de forma importante en función de las distintas opciones para cada factor de riesgo. No obstante, en un análisis descriptivo no recoge el impacto de cada factor de riesgo manteniendo invariable el resto de los factores algo que sí podremos hacer a través de los modelos GLM, concretamente el modelo de regresión de Poisson, el modelo de regresión Binomial Negativo y los modelos de regresión cero inflado parta estas mismas distribuciones.

Antes de aplicar los distintos modelos se realiza un análisis de sobredispersión de los datos ya que tal y como se ha detallado en el presente trabajo la regresión de Poisson parte de la premisa de equidispersión en los datos. Al analizar la dispersión de la base de datos se observa como si bien los datos no llegan a ser equidispersos la sobredispersión observada es muy ligera, lo cual permite emplear el modelo de regresión de Poisson. No sucederá esto, en cambio en el modelo de regresión Binomial Negativo cero inflado cuyo

modelo no converge. Esto sucede debido a que, una vez inflados los ceros, los datos dejan de presentar sobredispersión.

Realizado este ejercicio previo se procede a aplicar los distintos modelos. En ellos se puede observar cómo:

- Las características del asegurado son de elevada importancia en la frecuencia siniestral ya que el género, edad y experiencia siniestral son de gran importancia:

    - Todos los modelos reflejan la importancia del genero en la frecuencia siniestral. Si bien la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres, no permite su uso en la gestión realizada por las compañías aseguradoras.

    - Los perfiles con menos de 35 años registran una frecuencia siniestral importante. En este caso, el modelo de regresión Poisson cero inflado además recoge una reducción significativa de la probabilidad de presentar 0 siniestros. Los perfiles de edades superiores a los 55 años, en cambio, no sacan conclusiones demasiado significativas. Esto podría deberse a una falta de volumen.

    - La bonificación por buen historial de conducción se hace determinante en todos los modelos observando como a mayor bonificación menor frecuencia siniestral.

- Las tipologías del vehículo no parecen tan determinantes como su antigüedad. En lo que se refiere a la antigüedad se desprende un punto relevante y es que en vehículos de 2 a 5 años aumenta la probabilidad de no tener siniestros, pero a su vez en caso de tenerlos aumenta la frecuencia siniestral respecto al vehículo nuevo. Esto vendría explicado porque estos vehículos presentan un buen mantenimiento, pero a su vez suelen tener cambios importantes de garantías incentivando la declaración antes de reducir las garantías fruto de la depreciación del vehículo. También es importante destacar como las frecuencias se reducen en antigüedades iguales o mayores a 16 años con alta probabilidad asociados a las reducciones de garantías que suelen producirse en vehículos tan antiguos.

Teniendo en consideración los outputs obtenidos se puede observar cómo en líneas generales la información obtenida por estos modelos podrá ser de gran relevancia para la compañía.

Adicionalmente, teniendo en cuenta el criterio AIC el modelo que mejor ajusta nuestros datos será el modelo de regresión de Poisson Cero Inflado, aunque el criterio BIC considera que es suficiente con el modelo de regresión de Poisson pues la mejora obtenida no es suficiente para incluir un mayor número de regresores.

Podemos concluir que dado que los datos presentan una sobredispersión leve:

- El modelo de regresión de Poisson parte de la premisa de equidispersión en los datos y, por tanto, no contempla sobredispersión.

- El modelo de regresión Binomial Negativo incluye un parámetro adicional y contempla sobredispersión, no obstante, la inclusión de un parámetro no estaría justificada ya que el modelo no se ajusta mucho mejor que Poisson ya que la sobredispersión es muy reducida.

- El modelo de regresión Poisson cero inflado recoge dicha sobredispersión a través de los ceros inflados siendo equidispersa después. Esto hace que el modelo sea capaz de coger dicha leve sobredispersión, si bien es un modelo que emplea un mayor número de regresores y podría considerarse que esto no está justificado para la mejora que ofrece.

- El modelo de regresión Binomial cero inflado, por su parte, todavía recoge mayor sobredispersión por lo que no es necesario y nos serviría con el modelo de regresión Poisson cero inflado o incluso con el modelo de regresión Binomial Negativo.

7. Referencias

Alemany, R., Ayuso, M. y Bolancé, C. (2021). Modelos lineales generalizados (tarificación a priori). (Material de la asignatura: Modelos estadísticos aplicados). Universidad de Barcelona.

Almoguera, P. y Roslan, R. (2017). Ni un coche más en Singapur. EL PAÍS. Versión del 18 de abril de 2021: https://elpais.com/internacional/2017/11/08/mundo_global/1510138675_326328.html (

COE Price for Commercial Vehicles Reaches New High. (2013). TODAYonline. Acceso el 18 de abril de 2021: https://www.todayonline.com/singapore/coe-price-commercial-vehicles-reaches-new-high

Frees, E. W. (2010). Regression Modeling with Actuarial and Financial Applications. [Dataset]. Acceso el 11 de marzo de 2021. https://instruction.bus.wisc.edu/jfrees/jfreesbooks/regression%20modeling/bookwebdec2010/CSVData/SingaporeAuto.csv

Frees, E. W. (2010). *Instructors' Manual for Regression Modeling with Actuarial and Financial Applications.* University of Cambridge.

Frees, E. W. (2020). *Loss Data Analytics* (pp.285- 308).

López Jiménez, JM. (2019). *Análisis de la siniestralidad vial mediante modelos de datos de conteo.* Universidad de Sevilla.

LTA Statistics & Publications. Vehicle Quota (COE). (2021). Acceso el 18 de abril de 2021. https://www.lta.gov.sg/content/dam/ltagov/who_we_are/statistics_and_publications/statistics/pdf/COEQuotaAllocationRV.pdf

Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres. Boletín Oficial del Estado, núm. 71, de 23 de marzo de 2007. https://www.boe.es/buscar/act.php?id=BOE-A-2007-6115

NCSS Statistical Software. Negative Binomial Regression, Chapter 326. Acceso el 28 de abril de 2021: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

R Documentation. (2021). AER (version 1.2-9) dispersiontest function. Acceso 13 de marzo de 2021. https://cran.r-project.org/web/packages/insuranceData/insuranceData.pdf

Sarrasí Vizcarra, FJ. (2019). Matemática del reaseguro (pp. 27) (Material de la asignatura: Reaseguro). Universidad de Barcelona.

Tiwari, A. (2020). Modeling Insurance Claim Frequency – The Startup Medium. Acceso el 27 de marzo de 2021: https://medium.com/swlh/modeling-insurance-claim-frequency-a776f3bf41dc

Trzesiok, M. y Wolny—Dominiak, A. (2015). Package ‘insuranceData’. Acceso el 20 de mayo de 2021. https://www.rdocumentation.org/packages/AER/versions/1.2-9/topics/dispersiontest

Vives Brosa, J. (2002). *El diagnóstico de sobredispersión en modelos de análisis de datos de recuento.* Universidad Autónoma de Barcelona.

Yoshika, K. (2014). Models for excess zeros using pscl package (Hurdle and zero-inflated regression models) and their interpretations. Acceso el 05 de mayo de 2021: https://rpubs.com/kaz_yos/pscl-2

Anexos

Script del código R

```
setwd("C:/ MÁSTER ACTUARIALES/TFM")

library(data.table)

library(tidyverse)

install.packages('psych')

library(psych)

install.packages('dplyr')

library(dplyr)

install.packages('knitr')

library(knitr)

install.packages("MASS")

library(MASS)

install.packages("pscl")

library(pscl)

Data <- fread("SingaporeAutov4.csv",header="auto")

Data<-as.data.frame(Data)

#######TRATAMIENTO DE LOS DATOS###############

#FILTRO BASE DE DATOS, selección columnas

Data_DF<-Data[,1:8]

VAgeCat<-Data[,13]

BBDD_Last<-cbind(Data_DF,VAgeCat)

BBDD_Last_DF<-as.data.frame(BBDD_Last)

attach(BBDD_Last_DF)

str(BBDD_Last_DF)

#FILTRO BASE DE DATOS

Group_VehicleType<-ifelse(BBDD_Last_DF$VehicleType=="A","A",
ifelse(BBDD_Last_DF$VehicleType=="G","G", ifelse(BBDD_Last_DF$VehicleType=="M","M","O")))

BBDD_Last_DF<-cbind(BBDD_Last_DF,Group_VehicleType)

attach(BBDD_Last_DF)

#CREAMOS VARIABLE SINIESTROS POR EXPUESTO

Clm_Count_Matriz<-as.matrix(BBDD_Last$Clm_Count)
```

```
Clm_Count_Matriz<-as.numeric(Clm_Count_Matriz)

Exp_weights_Matriz<-as.matrix(BBDD_Last$Exp_weights)

xp_weights_Matriz<-as.numeric(Exp_weights_Matriz)

Clm_Exp<-Clm_Count_Matriz / Exp_weights_Matriz

BBDD_Last<-cbind(BBDD_Last_DF,Clm_Exp)

BBDD_Last<-as.data.frame(BBDD_Last)

rm(Clm_Exp)


sum(BBDD_Last$Clm_Exp)

mean(BBDD_Last$Clm_Exp)

median(BBDD_Last$Clm_Exp)

sd(BBDD_Last$Clm_Exp)

var(BBDD_Last$Clm_Exp)

skew(BBDD_Last$Clm_Exp)

kurtosi(BBDD_Last$Clm_Exp)

quantile(BBDD_Last$Clm_Exp)

quantile(BBDD_Last$Clm_Exp, probs = 0.95)

#######FACTORES DE RIESGO###############

#1.Descripción Género

ggplot(data = BBDD_Last)+ geom_bar(colour="skyblue4", fill="lightskyblue1", mapping = aes(x=SexInsured))+ xlab("Género")+ylab("Nº de observaciones")+ggtitle("Género del tomador")+ theme(plot.title = element_text(color="skyblue4", size=12, face="bold.italic", hjust = 0.5),axis.title.x = element_text(color="black", size=10, face="bold"),axis.title.y = element_text(color="black", size=10, face="bold"))

BBDD_Last$SexInsured<-factor(BBDD_Last$SexInsured)


sum(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

median(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

var(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

skew(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'])
```

```
quantile(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='F'], probs = 0.95)

sum(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

median(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

var(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

skew(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$SexInsured=='M'], probs = 0.95)

#2.Descripción Variable Tipo Vehiculo

ggplot(data = BBDD_Last)+geom_bar(colour="skyblue4",fill="lightskyblue1", mapping = aes(x=Group_VehicleType))+xlab("Tipo Vehículo")+ylab("Nº de observaciones")+ggtitle("Tipo de Vehículo")+theme( plot.title = element_text(color="skyblue4", size=12, face="bold.italic", hjust = 0.5),axis.title.x = element_text(color="black", size=10, face="bold"),axis.title.y = element_text(color="black", size=10, face="bold"))


sum(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='A'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])
```

```
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='G'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='M'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Group_VehicleType=='O'])

#3.Descripción Variable Uso

Uso<-ifelse(BBDD_Last$PC==1,"Privado","No Privado")

BBDD_Last<-cbind(BBDD_Last,Uso)

attach(BBDD_Last)

ggplot(data = BBDD_Last)+ geom_bar(colour="skyblue4",fill="lightskyblue1", mapping = aes(x=Uso))+
xlab("Uso")+ ylab("Nº de observaciones")+ ggtitle("Uso de Vehículo")+ theme(plot.title =
element_text(color="skyblue4", size=12, face="bold.italic", hjust = 0.5),axis.title.x =
element_text(color="black", size=10, face="bold"), axis.title.y = element_text(color="black", size=10,
face="bold") )


sum(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])
```

```
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='No Privado'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

median(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

sd(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

var(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'], probs = 0.95)

skew(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Uso=='Privado'])

#4.Descripción Variable NCD

attach(BBDD_Last)

BBDD_Last$NCD<-factor(BBDD_Last$NCD, levels=c("0", "10", "20", "30", "40", "50"))

ggplot(data = BBDD_Last)+geom_bar(colour="skyblue4",fill="lightskyblue1", mapping = aes(x=NCD))+
xlab("Bonificación")+ ylab("Nº de observaciones")+ ggtitle("Bonificación")+ theme(plot.title =
element_text(color="skyblue4", size=12, face="bold.italic", hjust = 0.5),axis.title.x =
element_text(color="black", size=10, face="bold"), axis.title.y = element_text(color="black", size=10,
face="bold") )

Rango_NCD<-ifelse(BBDD_Last$NCD=="0","Nivel 0 (0)",ifelse(BBDD_Last$NCD=="10","Nivel 1
(10)", ifelse(BBDD_Last$NCD=="20","Nivel 2(20-30)",ifelse(BBDD_Last$NCD=="30","Nivel 2(20-
30)","Nivel 3(40-50)")))))

BBDD_Last<-cbind(BBDD_Last,Rango_NCD)

BBDD_Last$Rango_NCD<-factor(BBDD_Last$Rango_NCD, levels=c("Nivel 0 (0)", "Nivel 1 (10)",
"Nivel 2(20-30)", "Nivel 3(40-50)"))

ggplot(data = BBDD_Last)+geom_bar(colour="skyblue4",fill="lightskyblue1", mapping =
aes(x=Rango_NCD))+ xlab("Bonificación")+ ylab("Nº de observaciones")+ ggtitle("Bonificación")+
theme( plot.title = element_text(color="skyblue4", size=12, face="bold.italic", hjust = 0.5), axis.title.x =
element_text(color="black", size=10, face="bold"), axis.title.y = element_text(color="black", size=10,
face="bold") )

Count<-rep(1,times=7483)

BBDD_Last<-cbind(BBDD_Last,Count)

xnames<-names(tapply(BBDD_Last$Count, BBDD_Last$Rango_NCD,sum))

plot(tapply(BBDD_Last$Count, BBDD_Last$Rango_NCD,sum),xlab="Bonificación", ylab="Nº de
observaciones" , type="b", xaxt="n", ylim=c(0,2500)), axis(1, at=1:length(xnames), labels=xnames)
```

```
sum(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
sum(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
sum(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
sum(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 0 (0)'], probs = 0.95)
```

```
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 1 (10)'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 2(20-30)'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Rango_NCD=='Nivel 3(40-50)'], probs = 0.95)

#5.Descripción Variable Edad

Rango_Edades<-ifelse(BBDD_Last$AgeCat=="0","=<21", ifelse(BBDD_Last$AgeCat=="2","[22,25]",
ifelse(BBDD_Last$AgeCat=="3","[26,35]", ifelse(BBDD_Last$AgeCat=="4","[36,45]",
ifelse(BBDD_Last$AgeCat=="5","[46,55]", ifelse(BBDD_Last$AgeCat=="6","[56,65]",">65"))))))

BBDD_Last<-cbind(BBDD_Last,Rango_Edades)

attach(BBDD_Last)

BBDD_Last$Rango_Edades<-factor(BBDD_Last$Rango_Edades, levels=c("=<21", "[22,25]", "[26,35]",
"[36,45]", "[46,55]", "[56,65]",">65"))

ggplot(data = BBDD_Last)+geom_bar(colour="skyblue4",fill="skyblue", mapping =
aes(x=Rango_Edades))+ xlab("Edad del Tomador")+ ylab("Nº de observaciones")

xnames<-names(tapply(BBDD_Last$Count, BBDD_Last$Rango_Edades,sum))

plot(tapply(BBDD_Last$Count, BBDD_Last$Rango_Edades,sum),xlab="Edad del Tomador", ylab="Nº
de observaciones" , type="b",xaxt="n", main="Edad del Tomador", col.main = "skyblue4", size.main=12,
face.main="bold.italic",col="skyblue4")

axis(1, at=1:length(xnames), labels=xnames)

Grupo_Rango_Edades<-
ifelse(BBDD_Last$AgeCat=="0","=<25",ifelse(BBDD_Last$AgeCat=="2","=<25",ifelse(BBDD_Last$
AgeCat=="3","[26,35]",ifelse(BBDD_Last$AgeCat=="4","[36,45]",ifelse(BBDD_Last$AgeCat=="5","[4
6,55]",">=56")))))

BBDD_Last<-cbind(BBDD_Last,Grupo_Rango_Edades)

attach(BBDD_Last)

BBDD_Last$Grupo_Rango_Edades<-factor(BBDD_Last$Grupo_Rango_Edades, levels=c("=<25",
"[26,35]", "[36,45]", "[46,55]", ">=56"))

xnames<-names(tapply(BBDD_Last$Count, BBDD_Last$Grupo_Rango_Edades,sum))

plot(tapply(BBDD_Last$Count, BBDD_Last$Grupo_Rango_Edades,sum),xlab="Edad del Tomador",
ylab="Nº de observaciones" , type="b",xaxt="n", main="Edad del Tomador", col.main = "skyblue4",
size.main=12, face.main="bold.italic", col="skyblue4", ylim = c(0,3000))

axis(1, at=1:length(xnames), labels=xnames)


sum(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])

sum(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])

mean(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
```

```
mean(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
mean(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
median(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
sd(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
var(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
skew(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'])
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'])
```

```
quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='=<25'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[26,35]'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[36,45]'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='[46,55]'], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$Grupo_Rango_Edades=='>=56'], probs = 0.95)

#6.Descripción Variable Antig.Veh.

Rango_Antig<-
ifelse(BBDD_Last$VAgeCat==0,"0",ifelse(BBDD_Last$VAgeCat==1,"1",ifelse(BBDD_Last$VAgeCat
==2,"2",ifelse(BBDD_Last$VAgeCat==3,"[3-5]", ifelse(BBDD_Last$VAgeCat==4,"[6-10]",
ifelse(BBDD_Last$VAgeCat==5,"[11-15]",">=16"))))))

BBDD_Last<-cbind(BBDD_Last,Rango_Antig)

attach(BBDD_Last)

BBDD_Last$Rango_Antig<-factor(BBDD_Last$Rango_Antig, levels=c("0", "1", "2", "[3-5]", "[6-10]",
"[11-15]",">=16"))

xnames<-names(tapply(BBDD_Last$Count, BBDD_Last$Rango_Antig,sum))

plot(tapply(BBDD_Last$Count, BBDD_Last$Rango_Antig,sum),xlab="Antigüedad del Vehículo",
ylab="Nº de observaciones" , type="b",xaxt="n", main="Antigüedad del Vehículo", col.main = "skyblue4",
size.main=12, face.main="bold.italic", col="skyblue4", ylim = c(0,3200))

axis(1, at=1:length(xnames), labels=xnames)


sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])

sum(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])

mean(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])
```

```
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])
median(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])
sd(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])
var(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])
skew(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])
```

```
kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])

kurtosi(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6])

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==0], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==1], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==2], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==3], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==4], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==5], probs = 0.95)

quantile(BBDD_Last$Clm_Exp[BBDD_Last$VAgeCat==6], probs = 0.95)

########### MODELO POISSON ###################

attach(BBDD_Last)

#######1. FACTOR GENERO######

# crear genero como elemento y caracter

SexIns = as.character(BBDD_Last$SexInsured)

# transformar a factor empleando de referencia F

Sex_RefF = relevel(factor(SexIns), ref = "F")

#######2. FACTOR TIPO VEH######

# crear tipo veh como elemento

Group_VehicleType_pois<-ifelse(BBDD_Last_DF$Group_VehicleType=="A",1,
ifelse(BBDD_Last_DF$Group_VehicleType=="G",2,ifelse(BBDD_Last_DF$Group_VehicleType=="M"
,3,4)))

#Transformar a factor y luego crear referencia

Group_VehType1 = relevel(factor(Group_VehicleType_pois), ref = 1)

#######3. FACTOR USO VEH######

# crear uso veh como elemento y caracter
```

```
UsoVeh = as.character(BBDD_Last$Uso)

#Transformar a factor y luego crear referencia

UsoVeh_RefP = relevel(factor(UsoVeh), ref = "Privado")

#######4. FACTOR NCD######

# crear uso veh como elemento

NCD_level_pois<-ifelse(BBDD_Last$Rango_NCD=="Nivel 0 (0)",1,
ifelse(BBDD_Last$Rango_NCD=="Nivel 1 (10)",2,ifelse(BBDD_Last$Rango_NCD=="Nivel 2(20-
30)",3,4)))

#Transformar a factor y luego crear referencia

NCD_level_Ref0 = relevel(factor(NCD_level_pois), ref = 1)

#######5. FACTOR Edad######

BBDD_Last$Grupo_Rango_Edades

Grupo_Edades_pois<-
ifelse(BBDD_Last$Grupo_Rango_Edades=="=<25",2,ifelse(BBDD_Last$Grupo_Rango_Edades=="[26,
35]",3,ifelse(BBDD_Last$Grupo_Rango_Edades=="[36,45]",1,ifelse(BBDD_Last$Grupo_Rango_Edade
s=="[46,55]",4,5))))

#Transformar a factor y luego crear referencia

Agecat_Ref3 = relevel( factor(Grupo_Edades_pois), ref = 1)

#######5. FACTOR Antig.Veh######

# crear uso veh como elemento

BBDD_Last$Rango_Antig

Grupo_Antig_pois<-ifelse(BBDD_Last$Rango_Antig=="0",1, ifelse(BBDD_Last$Rango_Antig=="1",2,
ifelse(BBDD_Last$Rango_Antig=="2",3,ifelse(BBDD_Last$Rango_Antig=="[3-5]",4,
ifelse(BBDD_Last$Rango_Antig=="[6-10]",5, ifelse(BBDD_Last$Rango_Antig=="[11-15]",6,7))))))

#Transformar a factor y luego crear referencia

VAgecat_Ref0 = relevel( factor(Grupo_Antig_pois), ref = 1 )

#BBDD_reestructuración

BBDD_Last$SexInsured = relevel(factor(SexIns), ref = "F")

BBDD_Last$Group_VehicleType = relevel(factor(Group_VehicleType_pois), ref = 1)

BBDD_Last$Uso = relevel(factor(UsoVeh), ref = "Privado")

BBDD_Last$Rango_NCD = relevel(factor(NCD_level_pois), ref = 1)

BBDD_Last$Grupo_Rango_Edades = relevel( factor(Grupo_Edades_pois), ref = 1)

BBDD_Last$Rango_Antig = relevel( factor(Grupo_Antig_pois), ref = 1 )

str(BBDD_Last)

####### Poisson reg model#######

BBDD_Last$LNWEIGHT<-as.numeric(BBDD_Last$LNWEIGHT)
```

```
LNWEIGHT<-BBDD_Last$LNWEIGHT

Pois_reg = glm(Clm_Count ~ SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig, offset = LNWEIGHT, poisson(link = log) )

#Pois_reg_ini = glm(Clm_Count ~ Sex_RefF + Group_VehType1 + UsoVeh_RefP + NCD_level_Ref0 + Agecat_Ref3 + VAgecat_Ref0, offset = LNWEIGHT, poisson(link = log) )

summary(Pois_reg)

#relativities

exp(Pois_reg$coefficients)

####### Binomial reg model#######

BN_reg = glm.nb(Clm_Count ~ SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig + offset(LNWEIGHT))

#BN_reg_ini = glm.nb(Clm_Count ~ Sex_RefF + Group_VehType1 + UsoVeh_RefP + NCD_level_Ref0 + Agecat_Ref3 + VAgecat_Ref0 + offset(LNWEIGHT))

summary(BN_reg)

exp(BN_reg$coefficients)

####### ZIP reg model#######

Pois_ZI = zeroinfl(Clm_Count ~ SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig | SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig,offset=log(Exp_weights), data = BBDD_Last, dist = "poisson", link = "logit")

summary(Pois_ZI)

######## ZIBN reg model#######

BN_ZI = zeroinfl(Clm_Count ~ SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig | SexInsured + Group_VehicleType + Uso + Rango_NCD + Grupo_Rango_Edades + Rango_Antig, offset=log(Exp_weights), data = BBDD_Last, dist = "negbin", link = "logit")

########Dispersión########

install.packages("AER")

library(AER)

dispersiontest(Pois_reg, trafo = 1)

########Contrastes########

AIC(Pois_reg)

AIC(BN_reg)

AIC(Pois_ZI)

BIC(Pois_reg)

BIC(BN_reg)

BIC(Pois_ZI)
```